Introduction to Topology

Pure and Applied

Colin Adams
Williams College

Robert Franzosa
University of Maine

Colin's Dedication:
To Alexa and Colton

Bob's Dedication:
Alla Famiglia

Contents

Preface

Topology is generally considered to be one of the three linchpins of modern abstract mathematics (along with analysis and algebra). In the early history of topology, results were primarily motivated by investigations of real-world problems. Then, after the formal foundation for topology was established in the first part of the twentieth century, the emphasis turned to its abstract development. However, within the past few decades there has been a significant increase in the applications of topology to fields as diverse as economics, engineering, chemistry, medicine, and cosmology.

Our goals in this book are twofold:

- To introduce topology as an important and fascinating mathematics discipline;

- To demonstrate the utility and significance of topological ideas in other areas of mathematics, in science, and in engineering.

A reader of this book will learn the basics of point-set topology and will be introduced to follow-up topics such as knots, manifolds, dynamical systems, fixed points, and topological graphs. Furthermore, the reader will learn how results from topology are used in applications that range from the atomic scale in chemistry to the astronomic scale in cosmology.

Intended Audience

A minimal background in formal mathematics is required to get started in this text. A student who has successfully completed an introductory course in abstract mathematics should have sufficient background. In Chapter 0 we provide an overview of the background material concerning sets, relations, functions, the real line, and Euclidean space needed for the remainder of the text. A mathematically mature student who has completed a calculus sequence and is capable of independently learning the Chapter 0 overview should be sufficiently prepared for this text. We view this text as appropriate for a one- or two-semester introduction to topology at the undergraduate level. It can also be used for an introductory graduate course.

How to Use This Text

The core introduction to point-set topology is found in the following sections in the text: 1.1–1.3, 2.1–2.3, 3.1–3.4, 4.1, 4.2, 5.1, 5.3, 6.1–6.4, 7.1–7.3.

We recommend covering these sections as a core for any course that uses this text. Much of the material in the remainder of the text uses many of the results from these sections.

Beyond the core sections, we have attempted to make the remainder of the topics in the text as independent of each other as possible. In a table below, we discuss and diagram the dependence between these topics.

For each of the topology topics beyond the core introduction, and for each of the applications, our approach is to provide a short, meaningful, and interesting introduction. Our hope is to spark the reader's interest in the topic or application and to plant a seed for further investigation.

A one-semester introductory topology course could cover the core material from Chapters 1–7, along with a few of the additional topics that appear in those and the remaining chapters of the text. A subsequent course could take a variety of forms. It could proceed through additional topics or applications in the text. It could be done in a seminar format where the students investigate the additional topics or applications and give presentations about them. It could have a focus on the applications of topology and mainly cover that material from the text. We also feel that the additional topics and applications lend themselves well to independent study by students and would provide a good introduction to material to be explored further for an honors or master's thesis.

Chapter and Section Interdependency

There are many pathways that can be followed through this text. Again, we recommend that the core introduction in Chapters 1–7 be covered as an intro-duction to point-set topology. On the next page there is a table of the book's chapters. It should help a reader or instructor when picking out topics to cover beyond the core introduction and when choosing an appropriate ordering of the topics. Here are some important points about the table:

- The sections corresponding to the core introductory material are circled.

- We do not address interdependency between Chapters 1–7, only interde-pendency between the sections within each of these chapters.

- The core material from Chapters 1–7 is prerequisite for the additional topics in Chapters 8–14, but none of the material from the non-core sections in Chapter 1–7 is required in Chapters 8–14 (with the exception of a few exercises, in which we make reference to the needed material). Therefore in the table, we do not address how Chapters 8–14 depend on Chapters 1–7.

- We address how Chapters 8–14 depend on each other, and within each of these chapters we also indicate how the sections depend on each other.

Proofs

In our development of topology as an abstract mathematics discipline, we provide the needed definitions, theorems, and proofs in their proper formality.

In the core introduction in Chapters 1–7, all of the theorems (with the exception of the Urysohn Metrization Theorem) are either proven or have proofs that are assigned as exercises. The proofs of these theorems primarily follow in a straightforward manner from the definitions and prior theorems. In the proofs that we present in these chapters, we have tried to provide good examples that readers can use as models when formulating their own proofs

Chapter 1: Topological Spaces	1.1 → 1.2 → 1.3 → 1.4
Chapter 2: Interior, Closure, and Boundary	2.1 → 2.2 → 2.3 → 2.4
Chapter 3: Creating New Topological Spaces	3.1 → 3.2 → 3.3 → 3.4 → 3.5
Chapter 4: Continuous Functions and Homeomorphisms	4.1 → 4.2 → 4.3
Chapter 5: Metric Spaces	5.1 → 5.3, with 5.1 → 5.2 and 5.3 → 5.4
Chapter 6: Connectedness	6.1 → 6.2 → 6.3 → 6.4 → 6.5
Chapter 7: Compactness	7.1 → 7.2 → 7.3 → 7.4, 7.5
Chapter 8: Dynamical Systems and Chaos • This chapter is independent of the other topics chapters.	8.1 → 8.2 → 8.3 → 8.4, 8.5
Chapter 9: Homotopy and Degree Theory • The proof of Theorem 9.14 uses the Tietze Extension Theorem which is established through supplementary exercises in Sections 4.1 and 7.3.	9.1 → 9.2 → 9.3, 9.4, 9.5, 9.6
Chapter 10: Fixed Point Theorems and Applications • The proof of the Brouwer Fixed Point Theorem in Section 10.1 uses the No Retraction Theorem from Section 9.2.	9.2 ↓ 10.1 → 10.2, 10.3 → 10.4
Chapter 11: Embeddings • The discussion of the Alexander Horned Sphere in Section 11.1 refers to simple connectivity, which is presented in Section 9.5. The full development of simple connectivity is not needed for this example—the definition and an intuitive understanding of the relevant results from Section 9.5 suffice. • The proof of the Jordan Curve Theorem in Section 11.2 requires the No Retraction Theorem and Theorem 9.14 from Section 9.2, as well as the Brouwer Fixed Point Theorem from Section 10.1. Note that, as mentioned under Chapter 9, the proof of Theorem 9.14 uses the Tietze Extension Theorem.	9.2 ↓ 10.1 ↓ 11.1 → 11.2, 11.3
Chapter 12: Knots • This chapter is independent of the other topics chapters.	12.1 → 12.2 → 12.3 → 12.4
Chapter 13: Graphs and Topology • This chapter is independent of the other topics chapters.	13.1 → 13.2, 13.3 → 13.4
Chapter 14: Manifolds and Cosmology • This chapter is independent of the other topics chapters, except that in the brief discussion about the Poincaré conjecture at the end of Section 14.3 we mention simple connectivity, a topic that is presented in Section 9.5.	14.1 → 14.2, 14.3 → 14.4 → 14.5

in assigned exercises. Reading and understanding proofs, and developing and writing them, are essential aspects of learning mathematics.

In Chapters 8–14, some of the theorems are presented without proofs, and some others are presented with proofs that exclude technical details. These omissions occur because the proofs or the details either require advanced tools beyond the scope of the text or are specialized for the setting and would take us too far off course to develop properly. We wish to keep the introductions to the topics in Chapters 8–14 relatively brief, and we wish to have them flow smoothly for the reader. In situations where proofs or details are excluded, we usually refer the reader to other sources where further information about the topic can be found.

Topology is a highly visual discipline, and therefore illustrations are often beneficial to the understanding of a topic under investigation. We occasionally use illustrations to communicate parts of an argument in a proof. This practice is common in topology, and there are advantages and disadvantages to doing so. An advantage is that it allows for a good flow of material in a book, a research work, or a professional presentation, reducing the details that need to be explicitly addressed. A disadvantage, particularly in an introductory text such as this, is that it can result in a perception that topology is mainly about pictures and is not as rigorous as other fields of mathematics. It is important to keep in mind that there is always formal mathematics behind the pictures, and, if need be, all of the necessary details can be provided to complete arguments that are represented by pictures.

Illustrations

As indicated previously, topology is very visual in nature. Therefore, we include many illustrations to aid the reader in understanding the material presented. While the content of the illustrations should be clear from the context, we have adopted a few conventions that we use in the illustrations as consistently as possible. We introduce these conventions here, using mathematical terms that are defined within the text.

General Topological Spaces: The topological space is the primary setting for the theory presented in the text. A rectangle with a black border and white interior usually represents a general topological space. For instance, Figure 1 depicts a set A in a general topological space X.
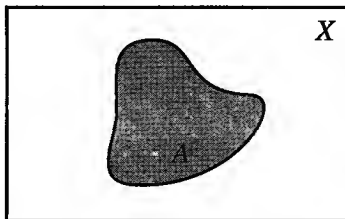
FIGURE 1: A set A in a topological space X.

The Plane: The plane is one of the main topological spaces used throughout the text. In illustrations, the plane is usually represented by a gray unbordered rectangle or parallelogram. For instance, Figure 2 depicts a set A in the plane—in one case viewing the plane directly, in the other case viewing the plane at an angle.
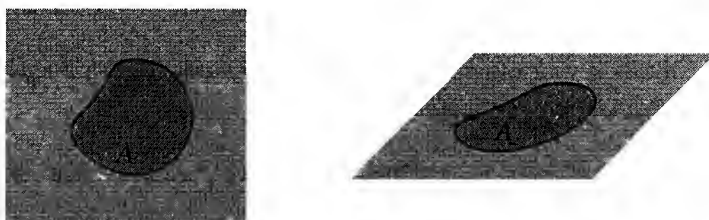
FIGURE 2: Two views of a set A in the plane.

Three-Space: Euclidean 3-space is also one of the main topological spaces used in the text. Often, it is left understood that an illustration depicts sets in 3-space. Occasionally, though, to emphasize that the setting for an illustration is 3-space, a set of three-dimensional coordinate axes is included in the illustration. For example, Figure 3(a) depicts a sphere in either no particular setting or a setting that is understood from the context, while Figure 3(b) depicts a sphere in 3-space.

(a)   (b)
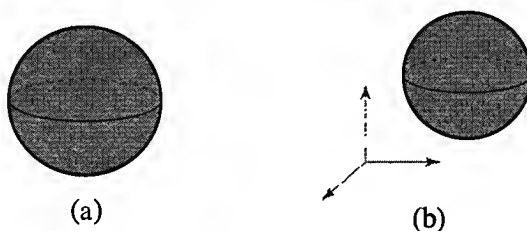
FIGURE 3: A sphere, and a sphere in three-space.

Open Sets and Excluded Boundary Parts: Topological spaces are collections of open sets. In illustrations, an open set is represented by a set that is bordered by a dashed curve. This convention reflects the fact that an open set does not include its boundary. Dashed curves are also used to indicate that a portion of the boundary of a set is not included in the set. In Figure 4, set A is an open set in a topological space X, set B is an open set in the plane, and set C is a square-shaped planar set that includes its left and right edges, but not its top and bottom ones.

Dashed curves are also used to indicate hidden or background curves in illustrations (as in Figure 3). Usually, these curves are thinner, are lighter gray, or have shorter dashes than the dashed curves used to represent open sets or

FIGURE 4: Sets A and B are open sets; set C does not include its top and bottom edges.

portions of boundaries that are excluded. Again, it should be clear from the context what such curves represent.

Excluded Points: Occasionally we consider sets that are obtained by removing points from other sets. In illustrations, the fact that a point is excluded from a set is usually depicted by using a small dashed circle as in Figure 5.
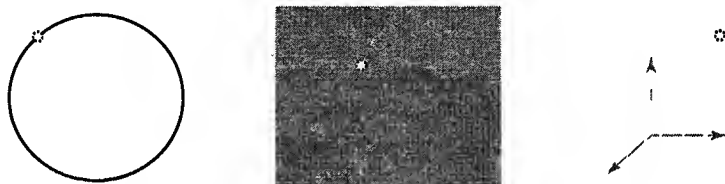
FIGURE 5: A circle with a point removed, a plane with a point removed, and 3-space with a point removed.

Acknowledgements

Many students have participated as audience and sounding board for various drafts of this text as we have been writing it. We are grateful to them for their input and feedback and for their patience dealing with a work in progress. We would particularly like to thank Kevin Roberge who worked through and critiqued many of the exercises in the text.

Different parts of the book have been reviewed during its development. We would like to thank all of the reviewers for their numerous helpful comments. The reviewers include Mark Brittenham, Robert Ghrist, Will Kazez, David Kraines, Krystyna Kuperberg, John McCleary, Dix H. Pettey, Daniel S. Silver, John Henry Steelman, and Gerard A. Venema.

We are very appreciative to Eric Franzosa for valuable assistance with the applications in biology and chemistry. We are also grateful to Peter Stadler, Dennis Garity, and Paul Franzosa for providing helpful input with other parts of the text material. A special thanks goes to Joey Shapiro for her input on the applications of topology to cosmology.

The original version of the text was written using Microsoft Word. We would like to thank Eric Franzosa for his assistance converting the text into LaTeX.

Chapter 0

Introduction

0.1 What is Topology and How is it Applied?

Topology is one of the most active areas in all of mathematics. Traditionally, it is considered one of the three main areas of pure mathematics (together with algebra and analysis). Recently, topology has also become an important component of applied mathematics, with many mathematicians and scientists employing concepts of topology to model and understand real-world structures and phenomena. In this book we introduce both aspects of topology: the traditional approach as an abstract discipline, and the emerging view as a valuable source of tools for applications. In the applied view, we see how topology impacts both real-world problems and a variety of results in other areas of mathematics.

Topology grew out of geometry, expanding on some of the ideas and loosening some of the structures appearing therein. The word topology, literally, means the study of position or location. Topology is the study of shapes, including their properties, deformations applied to them, mappings between them, and configurations composed of them.

Topology is often described as rubber-sheet geometry. In traditional geometry, objects such as circles, triangles, planes, and polyhedra are considered rigid, with well-defined distances between points and well-defined angles between edges or faces. But in topology, distances and angles are irrelevant. We treat objects as if they are made of rubber, capable of being deformed. We allow objects to be bent, twisted, stretched, shrunk, or otherwise deformed from one to another, but we do not allow the objects to be ripped apart. In Figure 0.1, we see four shapes that are very different from a geometric perspective, but are considered equivalent in topology. Any one of the four, if made of rubber, can be deformed to each of the others.
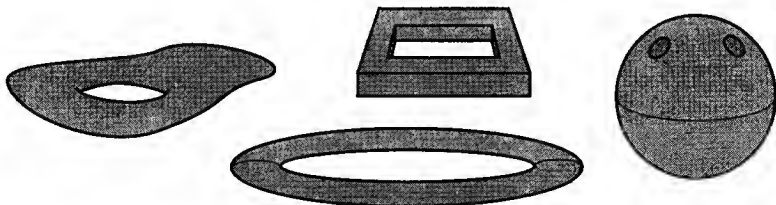
FIGURE 0.1: To a topologist these four objects are equivalent.

In Figure 0.2, we see two objects—the torus and the sphere—that are topologically distinct. We cannot deform a sphere into a torus in any permitted topological manner, and therefore they are not equivalent from a topological perspective.
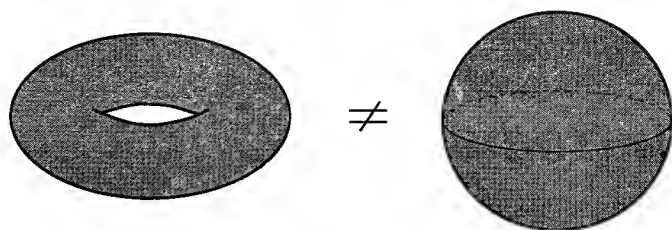
FIGURE 0.2: The torus and sphere are not topologically equivalent.

It is often said that a topologist cannot distinguish between a doughnut and a coffee cup. The point is that in topology a coffee cup can be deformed into the shape of a doughnut. (See Figure 0.3.) These objects are equivalent as far as topology is concerned.
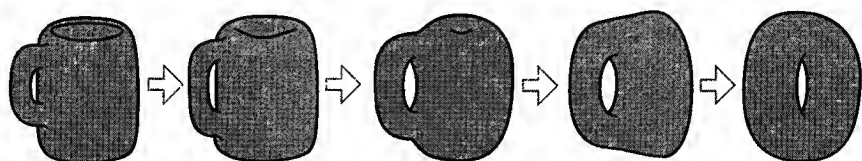
FIGURE 0.3: A coffee cup and doughnut are topologically equivalent.

You might wonder how this rubber-sheet geometry could be useful in real-world applications. The idea is very simple. Often, in a specific situation, the properties that are significant are those that are preserved when we treat an object as deformable, as opposed to those that are preserved when we treat that object as rigid. A topologist cannot distinguish between a doughnut and coffee cup, but what a topologist can do—and we will see how important this is throughout the text—is identify and use properties that these shapes have in common.

Let us take a quick look at a few topics from topology and the role they play in some of the applications that we present in this text. All of the ideas that we introduce here are developed more fully in the chapters that follow.

Topological Spaces and Phenotype Spaces: The objects that we study in topology are called topological spaces. These are sets of points on which a notion of proximity between points is established by specifying a collection of subsets called open sets. The line, the circle, the plane, the sphere, the torus, and the Möbius band are all examples of topological spaces. (See Figure 0.4.) In Chapters 1 and 3 we will see how these topological spaces and their open sets are specifically defined, and we will see how to construct topological spaces in a variety of settings and situations.

The concepts of genotype and phenotype are important in evolutionary biology. Every living organism is the physical realization (phenotype) of internally coded, inheritable information (genotype). Evolutionary change from one phenotype to another occurs by changes, called mutations, in their corresponding genotypes. We would like to have a notion of how close one
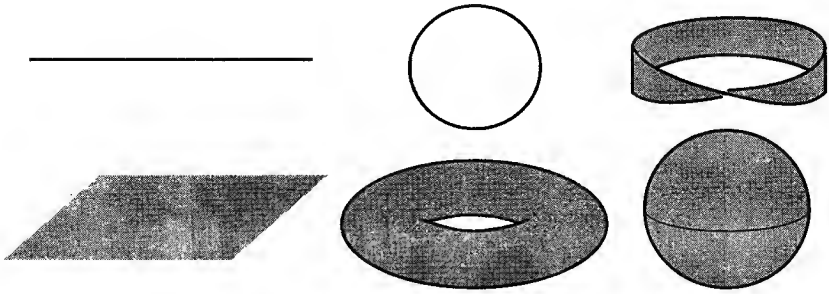
FIGURE 0.4: A variety of topological spaces.

phenotype is to another, reflecting how likely it is that a genotype mutation transforms one phenotype to the other. In Section 1.4 we describe how molecular biologists establish a notion of evolutionary proximity by constructing a topological space from a set of phenotypes.

Interior and Boundary, and Region Relationships in a Geographic Information System: Given a set in a topological space, two important sets that we associate to it are its interior and boundary. For a set A in a space X, we intuitively think of the interior of A as the points that are surrounded by nearby points in A, and we think of the boundary of A as the points that lie arbitrarily close to both A and the set of points that lie outside of A. (See Figure 0.5.) We formally define interior and boundary in terms of open sets in Chapter 2.
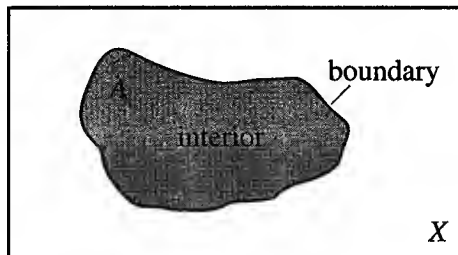
FIGURE 0.5: The interior and boundary of the set A.

A geographic information system (GIS) is a computer system that stores and manipulates data that are geographic in nature. A GIS must be able to address queries such as, "Does the planned Coltonian landfill overlap the Alexandria conservation zone?" In order to do so, the GIS needs to be able to distinguish a spectrum of relationships between land regions and needs to have a means for determining the particular relationship satisfied by a given pair of regions. In Section 2.4 we present a model for distinguishing such relationships based on the interiors and boundaries of sets. For example, if two sets A and B intersect only in their boundaries, then we say that A and B "touch" each other. In Figure 0.6 we illustrate that relationship and others that we identify in the model we present.

A is disjoint from B	A touches B	A equals B
A is within B	A contains B	A overlaps B
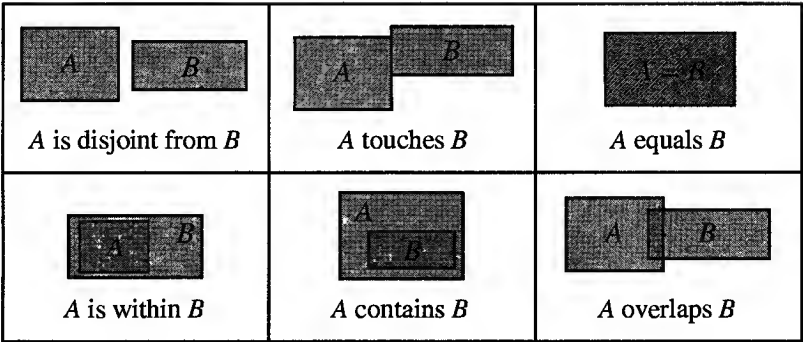
FIGURE 0.6: Possible relationships between sets A and B.

Manifolds and Cosmology: An n-manifold is a topological space that locally resembles n-dimensional Euclidean space. For example, a 1-manifold locally resembles a line, a 2-manifold locally resembles a plane, a 3-manifold locally resembles 3-space, and so on.

A 2-manifold is also called a surface. The sphere and torus are examples of surfaces. Each point in a surface lies in an open set that is topologically equivalent to an open set in the plane. (See Figure 0.7.)
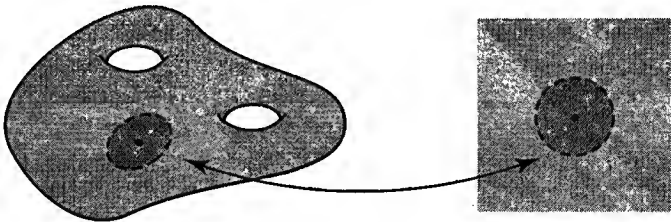
FIGURE 0.7: Locally a surface and a plane appear the same.

Inhabitants of surfaces might believe that they actually live in a plane because what they see locally appears the same as if they were on a plane. (This notion was exploited in the entertaining nineteenth-century book, *Flatland: A Romance of Many Dimensions*, by Edwin Abbott.) However, if the surface inhabitants studied properties of their surface world, they might be able to deduce its overall shape and therefore be able to distinguish it from a plane.

In a similar way, it is natural to feel that the universe in which we live is Euclidean 3-space because that is how we perceive it locally. Cosmologists try to determine the exact shape of the universe by studying properties that reveal aspects of its overall structure. In Chapter 14 we examine 3-manifolds, and we discuss approaches cosmologists are taking to determine which 3-manifold models the shape of our universe.

Embeddings, Knots, and DNA: An embedding is a function that takes one topological space and places a copy of it within another topological space. In Figure 0.8 we illustrate the results of two embeddings of a circle in 3-space.
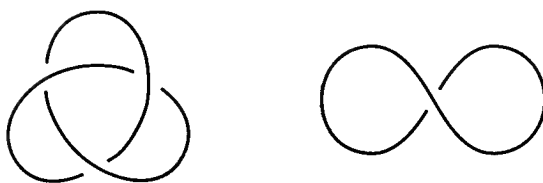
FIGURE 0.8: Embeddings of the circle in 3-space.

These embeddings appear distinctly different. On the left, the embedded circle is knotted, while on the right it is not. The study of embeddings of the circle in 3-space is called knot theory. Determining the different types of knots and how they are related to each other is an important aspect of this area of topology.

In knot theory, one of the fundamental operations that we perform on knots is a crossing change. As in Figure 0.9, we switch the crossing so that the understrand becomes the overstrand and vice versa. It is natural to ask what effect this operation has on a knot.
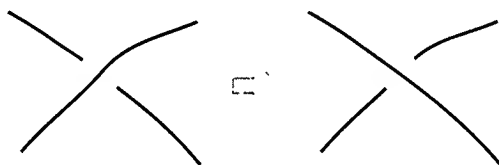
FIGURE 0.9: The crossing change operation.

We see this same operation occurring at the molecular level within the nucleus of the cell. Deoxyribonucleic acid (DNA) is a long, thin molecule made up of millions of atoms. This long strand is stuffed inside the nucleus of the cell, a packing comparable to placing 200 kilometers of tangled fishing line inside a basketball.

In order for essential life processes to work, the biological machinery of the cell must be free to access and manipulate the DNA molecule. The cell's ability to efficiently untangle DNA is therefore crucial to its survival. Inside the nucleus of the cell there are molecules, called enzymes, that act as biological tools. Some of these enzymes make crossing changes in the DNA, allowing it to untangle. An enzyme cuts the DNA strand at a place where it crosses over itself and then reattaches it so that the crossing is of the opposite type. Recently, new chemotherapy agents have been introduced that inhibit the enzymes from acting, thereby preventing cancerous DNA from re-creating itself. We discuss knot theory and this application in Chapter 12.

Fixed Points and Economics: A function that maps a topological space to itself has a fixed point if there is a particular point in the space that is mapped to itself by the function. For example, point p in Figure 0.10 is a fixed point of the function f illustrated there.

Fixed-point theory is an important area of topology; it involves the investigation of questions such as "Which functions mapping a topological space to
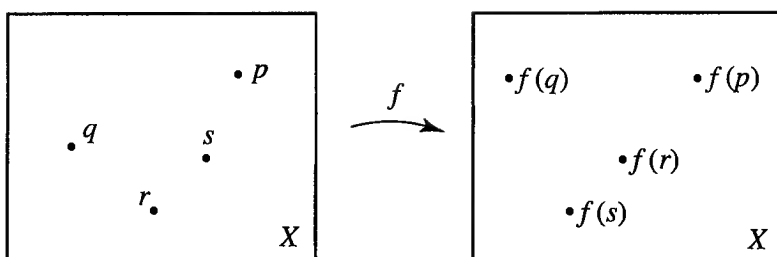
FIGURE 0.10: The function f has a fixed point at p.

itself have a fixed point?" and "Which topological spaces are such that every continuous function mapping the space to itself has a fixed point?" The most well-known result in this area is the Brouwer Fixed Point Theorem. It asserts that every continuous function on a closed n-dimensional ball has a fixed point. For instance, it says that in dimension one every continuous function from a closed interval to itself has a fixed point, and, in dimension two, every continuous function from a disk to itself has a fixed point. (See Figure 0.11.) We prove these one- and two-dimensional results in Chapters 6 and 10, respectively.
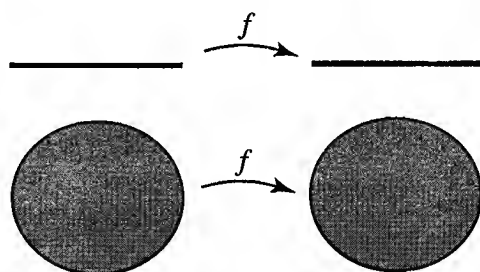
FIGURE 0.11: Continuous functions mapping the interval to the interval or the disk to the disk must have a fixed point.

In a model economic system there are consumers and goods and a set of variables associated with them, including the supply of the goods, the price of the goods, and the demand for the goods. In general, the values of the variables change in the near future, depending on their values at the present time. For example, a current short supply of applied topology texts might result in a higher price for them in the near future. Economists are interested in knowing if an economic system can be in equilibrium, a state where the consumers are appropriately satisfied and there are an unchanging supply, price, and demand associated with each of the goods. In Section 10.2 we apply the Brouwer Fixed Point Theorem in such an economic model and demonstrate that equilibrium states are possible. .

The conclusion of the Brouwer Fixed Point Theorem is an existence result. It asserts the existence of a fixed point, but it does not indicate where it is located. This is the nature of many of the results encountered in topology and

employed in applications. It is a demonstration of the qualitative nature of the conclusions derived using topological tools.

Since quantities such as distance and angle measure are irrelevant in topology, we do not expect to extract quantitative results in applications of topology. In general, topology aids in the qualitative analysis of a system. It helps identify possibilities or impossibilities that arise as a result of underlying structures.

There are three important components to an application of topology:

(i) Identification of the pertinent topological structures and relationships.

(ii) Proper definition of the topological concepts and rigorous verification of the topological results.

(iii) Application of the topological concepts and results to draw conclusions.

As a result, throughout the book we will have a threefold view of topology: a broad view, a theoretical view, and an applied view.

In the broad view, we intuitively regard topology as a general study of shapes and their properties. These are the topological aspects of applied problems that we identify and examine as we study real-world systems and draw conclusions about them.

The theoretical view of topology is the classic view of topology as an abstract mathematics discipline. We build this theory from its foundation so that we can properly define the concepts and structures that we investigate in the broad view and so that we have an appropriate theoretical basis to support the tools and relationships that we employ in the applied view.

Finally, in the applied view we study applications of topology involving the relationships and results that we develop. The examples discussed previously provide an initial look at the types of real-world applications we will consider. Furthermore, we examine how topological concepts help to define and reveal important structures and relationships in other areas of mathematics. For example, we identify topological properties underlying the Fundamental Theorem of Calculus in Chapter 7, we use topological concepts to define chaos in dynamical systems in Chapter 8, and we use topological results to prove the Fundamental Theorem of Algebra in Chapter 9.

0.2   A Glimpse at the History

Topology is popularly considered to have begun with Leonhard Euler's (1707–1783) solution to the famous Königsberg bridges problem. Let us take a quick look at this problem.

In the eighteenth century, the river Pregel flowed through the city of Königsberg, in Prussia (now Kaliningrad, in Russia), dividing it into four separate regions. There were seven bridges that crossed the river and connected the regions as illustrated in Figure 0.12.

A favored pastime was to take a bridge-crossing stroll through Königsberg. People asked, "Can you take a stroll through the city, crossing each bridge exactly once?" Curiously, no one was able to find a way to do so.
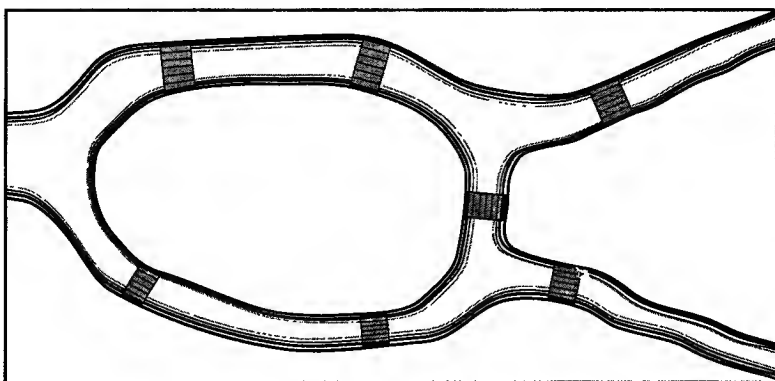
FIGURE 0.12: The Königsberg bridges. (Adapted from an illustration in [New].)

The problem caught the attention of Leonhard Euler, who realized that it involved a new mathematical approach, called "the geometry of position." He wrote:

> Recently, there was announced a problem which while it certainly seemed to belong to geometry, was nevertheless so designed that it did not call for the determination of magnitude, nor could it be solved by quantitative calculation; consequently I did not hesitate to assign it to the geometry of position, especially since the solution required only the consideration of position, calculation being of no use. [New]

Euler analyzed the problem and proved that it was impossible to accomplish the desired bridge-crossing walk, given the configuration of regions and bridges. We examine this problem further in Section 13.1.

Following the initial work of Euler, a number of prominent mathematicians made valuable contributions to the geometry of position over the next century and a half. These included Carl Friedrich Gauss (1777–1855), August Ferdinand Möbius (1790–1868), Johann Listing (1808–1882), Bernhard Riemann (1826–1866), Felix Klein (1849–1925), and Henri Poincaré (1854–1912).

The term "topology" first appeared in the title of Listing's paper, "Vorstudien zur Topologie" in 1847, but the term was not commonly used and the discipline was not formally defined until several decades later. "Analysis situs" was the expression primarily used for this area of geometry, and in the introduction to his 1895 paper "Analysis situs," Poincaré wrote of the geometry-of-position philosophy:

> The proportions of the figures might be grossly altered, but their elements must not be interchanged and must conserve their relative situation. In other terms, one does not worry about quantitative properties, but one must respect the qualitative properties, that is to say precisely those which are the concerns of Analysis Situs. [Sar]

Many nineteenth-century efforts in the geometry of position were motivated by applied problems, including James Clerk Maxwell's (1831–1879) and Peter Guthrie Tait's (1831–1901) work on knots (arising from investigations in chemistry), Gustav Kirchoff's (1824–1887) study of electrical networks, and Poincaré's analysis of celestial mechanics.

In the late nineteenth- and early twentieth-century there were numerous contributions to the growing discipline that would soon become the field of topology. L. E. J. Brouwer (1881–1966), Georg Cantor (1845–1918), Maurice Fréchet (1878–1973), Felix Hausdorff (1868–1962), Poincaré, Frigyes Riesz (1880–1956), and Hermann Weyl (1885–1955) were some of the mathematicians involved. Hausdorff's 1914 book *Grundzüge der Mengenlehre* (Fundamentals of Set Theory) introduced an axiomatic foundation for topological spaces and thereby initiated the general study of topology as an abstract mathematics discipline.

Throughout most of the twentieth century, topology developed primarily as a branch of abstract mathematics. However, in recent years there has been dramatic growth in the applications of topology. We explore topology from both of these perspectives in the chapters that follow.

In the remaining sections of this chapter, we review some basic topics involving sets, the real number system, Euclidean space, relations, and functions. These topics play an important role throughout the text. More detailed developments of this material can be found in introductory abstract mathematics texts such as [Blo] and [Hum].

0.3 Sets and Operations on Them

In this section we consider sets, relationships between them, and operations on them.

Throughout the text we employ standard set-builder notation for defining sets. We use the term **elements** to refer to members of sets, and we write $x \in A$ to mean x is an element of set A. Most of the sets that we encounter in topology are considered to be sets of points, so in those circumstances we use elements and points interchangeably. The set containing no elements is called the **empty set**; we denote it by $\varnothing$.

We denote the real line or the set of real numbers by $\mathbb{R}$. Important subsets of $\mathbb{R}$ that we employ are the positive real numbers $\mathbb{R}_+$, the integers $\mathbb{Z}$, the positive integers $\mathbb{Z}_+$, and the rational numbers $\mathbb{Q}$.

We assume all of the basic algebraic and order properties of the real number system.

One such property states that between every pair of real numbers there is an irrational number and a rational number. This property is useful in exploring the topology of the line. It follows that arbitrarily close to every real number there is a rational number. Consequently the set of rational numbers is said to be dense in $\mathbb{R}$. (See Section 2.1.) Similarly, the set of irrational numbers is dense in $\mathbb{R}$.

Another important property is that every real number is an element of an interval of the form $[n, n + 1]$ where n is an integer.

A fundamental defining property of the real number system is the **least upper bound property**: Every subset of $\mathbb{R}$ that is bounded from above has a least upper bound. Equivalent to this property is the **greatest lower bound property**: Every subset of $\mathbb{R}$ that is bounded from below has a greatest lower bound. We denote the least upper bound and greatest lower bound of a set A by $\text{lub}(A)$ and $\text{glb}(A)$, respectively.

We use the following standard notation for intervals in $\mathbb{R}$:

(a,b)	$=$	$\{x \in \mathbb{R} \mid a < x < b\}$	$[a,b]$	$=$	$\{x \in \mathbb{R} \mid a \le x \le b\}$
$[a,b)$	$=$	$\{x \in \mathbb{R} \mid a \le x < b\}$	$(a,b]$	$=$	$\{x \in \mathbb{R} \mid a < x \le b\}$
(a,∞)	$=$	$\{x \in \mathbb{R} \mid a < x\}$	$[a,\infty)$	$=$	$\{x \in \mathbb{R} \mid a \le x\}$
$(-\infty,b)$	$=$	$\{x \in \mathbb{R} \mid x < b\}$	$(-\infty,b]$	$=$	$\{x \in \mathbb{R} \mid x \le b\}$

The intervals (a,b), $(-\infty, b)$, and (a, ∞) are called **open intervals**; the intervals $[a,b]$, $(-\infty, b]$, and $[a, \infty)$ are called **closed intervals**; and the intervals $[a, b)$ and $(a, b]$ are called **half-open intervals**. Intervals of the form $[a, b]$ are also referred to as **closed bounded intervals**.

For sets A and B, we say that A **is a subset of** B, denoted $A \subset B$, if every element of A is also an element of B. Sets A and B are **equal**, written $A = B$, if both $A \subset B$ and $B \subset A$. If $A \subset B$ but A does not equal B, then we say that A **is a proper subset of** B.

We frequently work with collections of sets, particularly indexed collections of sets, defined as follows:

DEFINITION 0.1. *Let D be a set, and assume that for each element d in D we have a set S_d. The collection C consisting of the sets S_d, one for each $d \in D$, is called an **indexed collection of sets** or a **collection of sets indexed by** D. We refer to D as the **indexing set** for the collection. We denote the collection C by $\{S_d\}_{d \in D}$ or just $\{S_d\}$ when the indexing set is clear or does not need to be specified.*

EXAMPLE 0.1. Let T be the set of towns in Arkansas, and for each $t \in T$, let B_t denote the set of barbers who run a business in town t. Note that two sets B_t and $B_{t'}$ have a barber in common if there is a barber who runs a business in both town t and town t'. The collection $\{B_t\}_{t \in T}$ is a collection of sets of barbers, with each set in the collection representing the barbers who run a business in a town in Arkansas.

A special case of an indexed collection of sets $\{S_d\}_{d \in D}$ occurs when the indexing set is a set of positive integers, $D = \{m, m + 1, \ldots, n\}$. In that case we also express the collection as $\{S_d\}_{d=m}^{n}$. We include here the possibility that $m = -\infty$ or $n = \infty$.

DEFINITION 0.2. *A collection of sets $\{S_d\}_{d=m}^{n}$ is said to be **nested** if $S_{d+1} \subset S_d$ for all d.*

EXAMPLE 0.2. The collection of intervals $[-\frac{1}{n}, \frac{1}{n}]$, with $n \in \mathbb{Z}_+$, is a nested collection of sets.

DEFINITION 0.3.

(i) *Let U and V be sets. The **union of U and V** is the set*

$$U \cup V = \{x \mid x \in U \text{ or } x \in V\},$$

*and the **intersection of U and V** is the set*

$$U \cap V = \{x \mid x \in U \text{ and } x \in V\}.$$

(ii) *Let $\{U_d\}_{d \in D}$ be an indexed collection of sets. Then the **union of the sets U_d** is the set*

$$\bigcup_{d \in D} U_d = \{x \mid x \in U_d \text{ for some } d \in D\},$$

*and the **intersection of the sets U_d** is the set*

$$\bigcap_{d \in D} U_d = \{x \mid x \in U_d \text{ for all } d \in D\}.$$

*We also denote these unions and intersections by $\bigcup U_d$ and $\bigcap U_d$, respectively. We refer to this type of union as an **arbitrary union** and this type of intersection as an **arbitrary intersection**.*

EXAMPLE 0.3. Let $\{B_t\}_{t \in T}$ be the collection of sets of barbers from Example 0.1. Then $\bigcup_{d \in D} U_d$ is the set of barbers who run a business in some town in Arkansas. And if $\bigcap_{d \in D} U_d$ is not the empty set, then each barber in this intersection runs a business in every town in Arkansas.

EXAMPLE 0.4. For each r in $\mathbb{R}_+$, let $U_r = [-r, r]$. Then $\bigcup_{r \in \mathbb{R}_+} U_r = \mathbb{R}$ and $\bigcap_{r \in \mathbb{R}_+} U_r = \{0\}$.

If the indexing set for $\{U_d\}_{d \in D}$ is a set of integers, $D = \{m, m+1, \ldots, n\}$, then we also express the union and intersection of the sets in the collection as $\bigcup_{d=m}^{n} U_d$ and $\bigcap_{d=m}^{n} U_d$, respectively. In the case where neither $m = -\infty$ nor $n = \infty$, we refer to these as a **finite union** and a **finite intersection**.

DEFINITION 0.4.

(i) *Let A and B be sets. If $A \cap B = \varnothing$, then we say that A and B are **disjoint**.*

(ii) *If $\{A_d\}_{d \in D}$ is a collection of sets, and every pair of sets in the collection is disjoint, then we say that the sets in the collection are **mutually disjoint**.*

The following lemma will be helpful to us when proving that a set is a union of a particular collection its subsets:

LEMMA 0.5. *The Union Lemma.* *Let X be a set and C be a collection of subsets of X. Assume that for each $x \in X$, there is a set A_x in C such that $x \in A_x$. Then $\bigcup_{x \in X} A_x = X$. (See Figure 0.13.)*
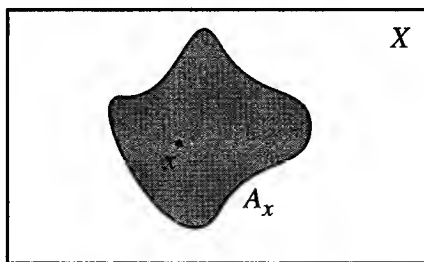
FIGURE 0.13: The set X is the union of the subsets A_x.

Proof. We prove that $\bigcup A_x \subset X$ and $X \subset \bigcup A_x$. First, since each A_x is a subset of X, it follows that $\bigcup A_x \subset X$.

Now suppose that $y \in X$. There exists A_y in C such that $y \in A_y$. Thus $y \in A_y \subset \bigcup A_x$. Since $y \in X$ implies that $y \in \bigcup A_x$, it follows that $X \subset \bigcup A_x$. Hence $\bigcup A_x = X$. ∎

EXAMPLE 0.5. Consider the collection of intervals in $\mathbb{R}$ given by $C = \{[s, s+1]\}_{s \in \mathbb{Z}}$. Every $x \in \mathbb{R}$ is an element of $[s, s+1]$ for some $s \in \mathbb{Z}$, and therefore the Union Lemma implies that $\mathbb{R}$ is the union of the intervals in C. (See Figure 0.14.)
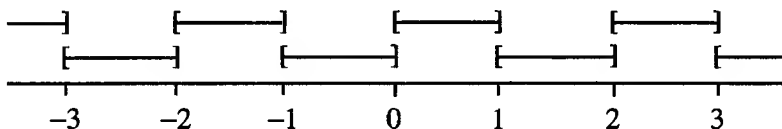
FIGURE 0.14: The set of real numbers is the union of all intervals of the form $[s, s+1]$ for $s \in \mathbb{Z}$.

DEFINITION 0.6. *Given two sets A and B, we define the **complement of B in A** to be the set $A - B = \{x \mid x \in A \text{ and } x \notin B\}$.*

The complement of B in A is a subset of A. For example, if $A \subset B$, then $A - B = \varnothing$, and if A and B are disjoint, then $A - B = A$. (See Figure 0.15.)

In subsequent chapters, we often work within a fixed set X. In that case, given a subset B of X, we say "the complement of B" instead of "the complement of B in X".

The complement of B in A is also called the **difference between A and B** and sometimes is denoted $A \backslash B$ in mathematics literature.
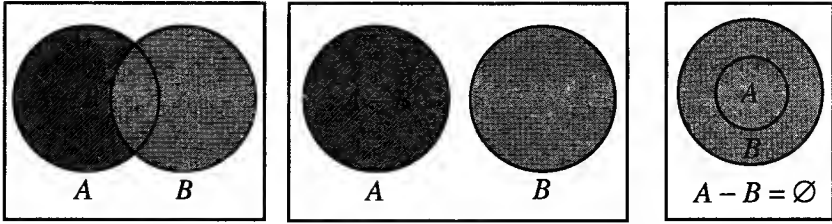
FIGURE 0.15: Various possibilities for the complement of B in A.

DEFINITION 0.7. *Given two sets A and B, we define the* **product** *of A and B to be the set*

$$A \times B = \{(a, b) \mid a \in A, \; b \in B\}.$$

The elements of $A \times B$ are ordered pairs of elements, one from A and one from B.

IMPORTANT NOTE: *The expression (a, b) is used to denote both an interval in $\mathbb{R}$ and an element in a product $A \times B$. It should be apparent from the context which meaning is intended.*

Notice that if A has m elements and B has n elements, then the product $A \times B$ has mn elements. We also work with finite products of sets, defined as follows:

DEFINITION 0.8. *The* **product** *of sets $A_1, A_2, \ldots , A_n$ is the set*

$$A_1 \times A_2 \times \ldots \times A_n = \{(a_1, a_2, \ldots , a_n) \mid a_i \in A_i, \; for \; i = 1, 2, \ldots , n\}.$$

The elements of the product are ordered n-tuples of elements from the sets A_i.

The following are some important relationships associated with sets and the aforementioned relations and operations. Throughout the text we use these and other similar relationships.

THEOREM 0.9. *For sets A, B, and C, the following laws hold:*

Distributive Laws:

$$A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$$
$$A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$$
$$A \times (B \cup C) = (A \times B) \cup (A \times C)$$
$$A \times (B \cap C) = (A \times B) \cap (A \times C)$$
$$A \times (B - C) = (A \times B) - (A \times C)$$

DeMorgan's Laws:

$$A - (B \cup C) = (A - B) \cap (A - C)$$
$$A - (B \cap C) = (A - B) \cup (A - C)$$

0.4   Euclidean Space

One of the most common sets used in topology is $\mathbb{R}^n$ or Euclidean n-space. **The plane**, denoted $\mathbb{R}^2$, is the set of ordered pairs of real numbers,

$$\mathbb{R}^2 = \{(x_1, x_2) \mid x_1, x_2 \in \mathbb{R}\}.$$

Thus $\mathbb{R}^2$ is the product, $\mathbb{R} \times \mathbb{R}$. (See Figure 0.16.)
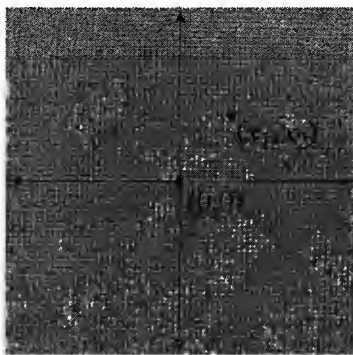
FIGURE 0.16: The plane $\mathbb{R}^2 = \{(x_1, x_2) \mid x_1, x_2 \in \mathbb{R}\}$.

In general, $\mathbb{R}^n$ is the product of n copies of the real line. It is the set of n-tuples of real numbers,

$$\mathbb{R}^n = \{(x_1, x_2, \dots, x_n) \mid x_1, x_2, \dots, x_n \in \mathbb{R}\}.$$

We call $\mathbb{R}^n$ **Euclidean n-space** or just n-**space** for short. The point $(0, 0, \dots, 0) \in \mathbb{R}^n$ is called **the origin** and we denote it by O.

The standard means for measuring distance in $\mathbb{R}^n$ makes use of the **Euclidean distance formula**, defined as follows: For $p = (p_1, \dots, p_n)$ and $q = (q_1, \dots, q_n)$, the distance between p and q is

$$d(p, q) = \sqrt{(p_1 - q_1)^2 + \dots + (p_n - q_n)^2},$$

as illustrated for $\mathbb{R}^2$ in Figure 0.17.

The Euclidean distance formula satisfies three important properties that make it what is known as a metric:

(i) For all $p, q \in \mathbb{R}^n$, $d(p, q) \geq 0$ and $d(p, q) = 0$ if and only if $p = q$.

(ii) For all $p, q \in \mathbb{R}^n$, $d(p, q) = d(q, p)$.

(iii) For all $p, q, r \in \mathbb{R}^n$, $d(p, r) \leq d(p, q) + d(q, r)$.

We discuss metrics and their relation to topology in Chapter 5. We regard each x in $\mathbb{R}^n$ as a point in n-space, but we can also regard it as a vector based at the origin in n-space. Either way, for each $x = (x_1, \dots, x_n)$ we define the
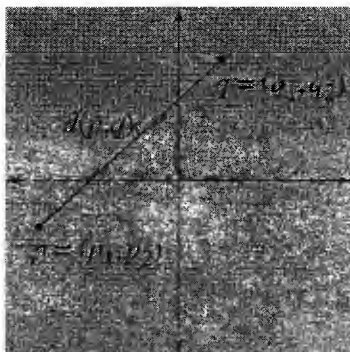
FIGURE 0.17: Measuring distance in $\mathbb{R}^2$.

norm of x by $|x| = \sqrt{x_1^2 + \ldots + x_n^2}$. We can think of the norm of x as either the distance from the origin to x or the length of x as a vector.

DEFINITION 0.10. *A set* $A \in \mathbb{R}^n$ *is called* **bounded** *if there exists* $b \in \mathbb{R}$ *such that* $|x| \leq b$ *for all* $x \in A$.

If a set $A \in \mathbb{R}^n$ is bounded, then there exists $d^* \in \mathbb{R}$ such that $d(p, q) \leq d^*$ for all $p, q \in A$. Thus, in a bounded set A in $\mathbb{R}^n$, the distance between all pairs of points in A is bounded from above.

DEFINITION 0.11. *A set* $A \subset \mathbb{R}^n$ *is called* **convex** *if for every* $p, q \in A$ *the line segment between* p *and* q *lies in* A.

For example, in Figure 0.18 sets A and B are convex subsets of the plane, but set C is not since there are pairs of points in C such that the line segments between the points do not lie in C.
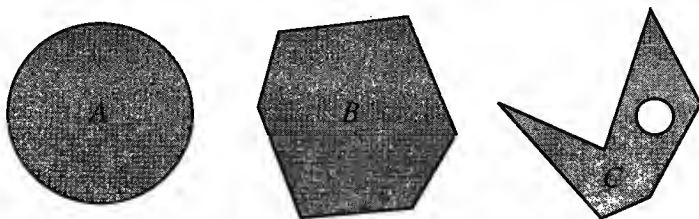
FIGURE 0.18: Sets A and B are convex in $\mathbb{R}^2$, but set C is not.

DEFINITION 0.12.

    (i) A **half space** *in* $\mathbb{R}^n$ *is a set of points* $(x_1, \ldots, x_n)$ *satisfying a linear inequality* $a_1 x_1 + \ldots + a_n x_n \leq b$ *where* b *and each* a_i *are real numbers and at least one of the* a_i *is nonzero.*

    (ii) A **polyhedron** *in* $\mathbb{R}^n$ *is a bounded subset of* $\mathbb{R}^n$ *that can be expressed as an intersection of half-spaces.*

In Figure 0.19 we show one-dimensional, two-dimensional, and three-dimensional polyhedra. A one-dimensional polyhedron is a line segment. A two-dimensional polyhedron is a polygon. Its perimeter is made up of vertices and edges. A three-dimensional polyhedron has a surface made up of polygonal faces that meet at edges and vertices.
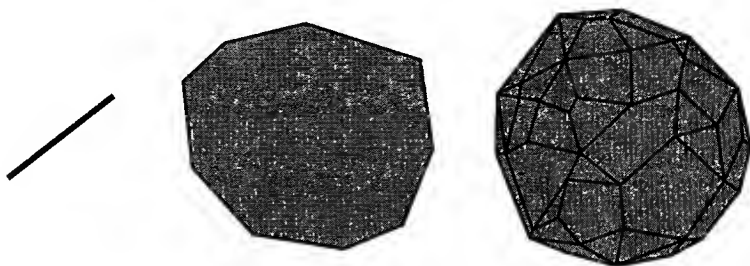
FIGURE 0.19: One-dimensional, two-dimensional, and three-dimensional polyhedra.

DEFINITION 0.13. *The n-sphere, denoted S^n, is the set of points that are distance 1 away from the origin in $\mathbb{R}^{n+1}$. Therefore,*

$$S^n = \{(x_1, \dots, x_{n+1}) \in \mathbb{R}^{n+1} \mid x_1^2 + \dots + x_{n+1}^2 = 1\}.$$

*The 1-sphere, S^1, is also called the **circle**, and the 2-sphere, S^2, is referred to simply as the **sphere**. (See Figure 0.20.)*

S^0 S^1 S^2
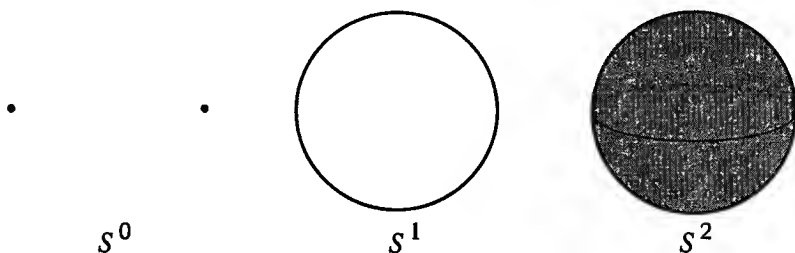
FIGURE 0.20: The 0-sphere, 1-sphere (circle), and 2-sphere.

If $x = (x_1, \dots, x_{n+1})$ is a point on the n-sphere, then the point $-x = (-x_1, \dots, -x_{n+1})$ is called the **antipode** of x or the point **antipodal** to x. Thus, the antipode of x is the point on the n-sphere that is opposite x, through the origin.

DEFINITION 0.14. *Define B^n, the **n-ball**, to be the set*

$$B^n = \{(x_1, \dots, x_n) \in \mathbb{R}^n \mid x_1^2 + \dots + x_n^2 \leq 1\}.$$

*The 2-ball is referred to as the **disk**. Further, define $\mathring{B}^n$, the **open n-ball**, to be the set*

$$\mathring{B}^n = \{(x_1, \ldots, x_n) \in \mathbb{R}^n \mid x_1^2 + \ldots + x_n^2 < 1\}.$$

*The open 2-ball is referred to as the **open disk**.*

0.5 Relations

DEFINITION 0.15. *Given sets X and Y,*

(i) *A **relation between X and Y** is a subset R of the product $X \times Y$. For a relation R, we say that x **is related** to y, denoted $x R y$, if $(x, y) \in R$.*

(ii) *A **relation on X** is a relation from X to X.*

An important type of relation used extensively in mathematics is the equivalence relation, defined as follows:

DEFINITION 0.16. *A relation $\sim$ on X is called an **equivalence relation** if the following properties hold for all a, b, c in X:*

(i) *$a \sim a$ (reflexivity),*

(ii) *$a \sim b$ implies $b \sim a$ (symmetry),*

(iii) *$a \sim b$ and $b \sim c$ imply $a \sim c$ (transitivity).*

EXAMPLE 0.6. Let D be the set of all dogs in Dubuque, Iowa, at this instant. Let $a \sim b$ if dog a has the same mother as dog b. We show that $\sim$ is an equivalence relation on D. Certainly each dog is related to itself, so reflexivity holds. Also, if Butterscotch has the same mother as Bingo, then Bingo has the same mother as Butterscotch. Thus symmetry holds. Finally, if Butterscotch and Bingo share a mother and Bingo and Rosco share a mother, then Butterscotch and Rosco also share a mother. Hence transitivity holds as well, and it follows that $\sim$ is an equivalence relation.

EXAMPLE 0.7. Once again, let D be the set of all dogs in Dubuque, Iowa, at this instant. But now define a relation L by $a L b$ if dog a has ever licked dog b. Although it is not a mathematical certainty, it is most likely true that reflexivity holds for this relation. Every dog has licked itself at one time or another. But the fact that Bucky licked Daisy does not mean that Daisy licked Bucky back, so symmetry does not necessarily hold. Similarly, Bucky having licked Daisy and Daisy having licked Zoe does not imply that Bucky ever licked Zoe. Thus transitivity need not apply either. Therefore L is a relation that is not necessarily an equivalence relation.

EXAMPLE 0.8. On $\mathbb{R}^2$ define $(p_1, p_2) \sim (q_1, q_2)$ if $p_1 + p_2 = q_1 + q_2$. We show that $\sim$ defines an equivalence relation.

(i) Since $p_1 + p_2 = p_1 + p_2$, it follows that $(p_1, p_2) \sim (p_1, p_2)$ for all (p_1, p_2) in $\mathbb{R}^2$.

(ii) If $(p_1, p_2) \sim (q_1, q_2)$ then $p_1 + p_2 = q_1 + q_2$, implying that $q_1 + q_2 = p_1 + p_2$, and therefore $(q_1, q_2) \sim (p_1, p_2)$. Thus for all $(p_1, p_2), (q_1, q_2)$ in $\mathbb{R}^2$, if $(p_1, p_2) \sim (q_1, q_2)$, then $(q_1, q_2) \sim (p_1, p_2)$.

(iii) Finally, assume $(p_1, p_2) \sim (q_1, q_2)$ and $(q_1, q_2) \sim (r_1, r_2)$. Then

$$p_1 + p_2 = q_1 + q_2 = r_1 + r_2,$$

implying that $(p_1, p_2) \sim (r_1, r_2)$. Hence for all $(p_1, p_2), (q_1, q_2)$, and (r_1, r_2) in $\mathbb{R}^2$, if $(p_1, p_2) \sim (q_1, q_2)$ and $(q_1, q_2) \sim (r_1, r_2)$, then $(p_1, p_2) \sim (r_1, r_2)$.

It follows that $\sim$ is an equivalence relation.

Now, given a point (a_1, a_2) in $\mathbb{R}^2$, can we identify all of the points that are equivalent to it? For (x_1, x_2) to be equivalent to (a_1, a_2), we must have $x_1 + x_2 = a_1 + a_2$. Thus all of the points (x_1, x_2) in the plane that are equivalent to (a_1, a_2) lie on the line $x_1 + x_2 = a_1 + a_2$. Furthermore, given any line $x_1 + x_2 = c$, every point on the line is equivalent to every other point on the line.

The collection of lines described in Example 0.8 partitions the plane into separate subsets. We formalize this idea in the following definition.

DEFINITION 0.17. *A partition of a set X is a collection of mutually disjoint subsets of X whose union is X.*

As in Example 0.8, an equivalence relation $\sim$ on a set X determines a partition of the set. For each $x \in X$ we define $[x] = \{p \in X \mid p \sim x\}$. We call $[x]$ **the equivalence class of x under** $\sim$; it is the set of all elements of X that are equivalent to x under the equivalence relation $\sim$. The collection of all equivalence classes under $\sim$ is a partition of X. For the equivalence relation in Example 0.8, the equivalence classes are the lines $x_1 + x_2 = c$ in $\mathbb{R}^2$, as illustrated in Figure 0.21.

We can also reverse this process. A partition of a set X defines an equivalence relation on X. We consider two elements of X to be related if they are contained in the same subset of the partition. This yields an equivalence relation and the resulting equivalence classes are the original sets in the partition.

EXAMPLE 0.9. Returning to Example 0.6, we immediately see that the equivalence relation defined there divides the population of dogs in Dubuque into subsets where each subset is the set of all dogs in Dubuque that are offspring of a particular mother.
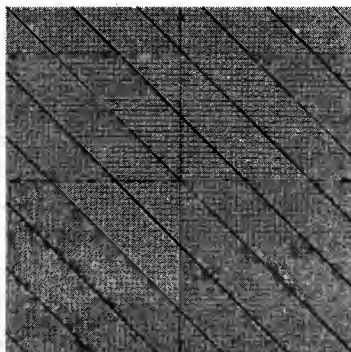
FIGURE 0.21: A partition of the plane into lines $x_1 + x_2 = c$.

0.6   *Functions*

As with other areas of mathematics, functions play an important role in topology. In this section we review basic definitions and properties related to functions.

DEFINITION 0.18. *Let X and Y be sets. A **function** f from X to Y is a relation between X and Y that associates to each x in X a unique y in Y. We write $f(x) = y$ and call y the **image of x under** f. The set X is called the **domain** of f, and the set Y is called the **range** of f.*

We use the notation $f : X \to Y$ to indicate that f is a function from X to Y. We also call a function $f : X \to Y$ a **mapping** and say that f **maps** X **to** Y. Further, if $f(x) = y$, then we say that f **maps** x **to** y.

DEFINITION 0.19. *For a function $f : X \to Y$ and a subset A of X, define the **image of** A **under** f to be the set*

$$f(A) = \{y \in Y \mid y = f(x) \text{ for some } x \in A\},$$

*as illustrated in Figure 0.22. The set $f(X)$ is the image of the domain; we also refer to it as the **image of** f.*
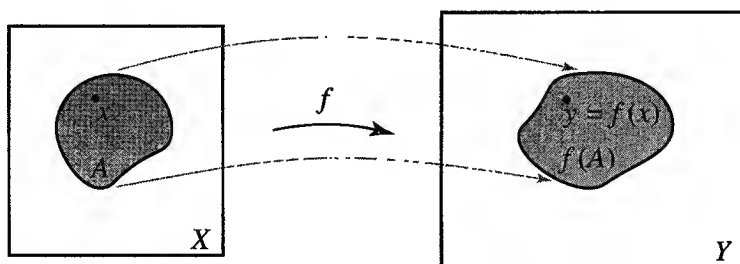
FIGURE 0.22: The image of A under f.

A function $f : X \to Y$ is called a **constant function** if its image consists of a single point $c \in Y$. In that case $f(x) = c$ for all $x \in X$. For a set X, the **identity on** X is the function $id_X : X \to X$, defined by $id_X(x) = x$ for all $x \in X$. The identity on X just maps each point x in X to itself.

DEFINITION 0.20. *Given $f : X \to Y$ and a point $y \in Y$, define $f^{-1}(y)$, the* **preimage of** *y, to be the set $\{x \in X \mid f(x) = y\}$. Furthermore, given a subset W of Y, define $f^{-1}(W)$, the* **preimage of** W, *to be the set $\{x \in X \mid f(x) \in W\}$.*

For a function $f : X \to Y$, the preimage of a point y is the set of points in X that f maps to y, and the preimage of a set W is the set of points in X that f maps to points in W.

EXAMPLE 0.10. If $f : \mathbb{R} \to \mathbb{R}$ is defined by $f(x) = x^2$, then $f^{-1}([0, 2]) = [-\sqrt{?}, \sqrt{2}]$ and $f^{-1}([1, 3]) = [-\sqrt{3}, -1] \cup [1, \sqrt{3}]$.

The following theorems state some basic relationships between images and preimages and operations on sets. We use these and other similar properties throughout the text.

THEOREM 0.21. *If $f : X \to Y$ is a function and A and B are subsets of X, then*

    (i) $f(A \cup B) = f(A) \cup f(B)$.

    (ii) $f(A \cap B) \subset f(A) \cap f(B)$.

    (iii) $f(A) - f(B) \subset f(A - B)$.

THEOREM 0.22. *If $f : X \to Y$ is a function and V and W are subsets of Y, then*

    (i) $f^{-1}(V \cup W) = f^{-1}(V) \cup f^{-1}(W)$.

    (ii) $f^{-1}(V \cap W) = f^{-1}(V) \cap f^{-1}(W)$.

    (iii) $f^{-1}(V - W) = f^{-1}(V) - f^{-1}(W)$.

DEFINITION 0.23.

    (i) A function $f : X \to Y$ is said to be **one-to-one** *or* **injective** *if $f(w) = f(x)$ implies $w = x$ for every $w, x \in X$. Thus f is injective if it maps distinct pairs of elements of X to distinct pairs of elements of Y.*

    (ii) A function $f : X \to Y$ is said to be **onto** *or* **surjective** *if $f(X) = Y$. That is, f is surjective if for every y in Y there is an x in X such that f maps x to y.*

(iii) *A function that is both injective and surjective is called* **bijective** *or a* **bijection**.

Being related by a bijection is the basic equivalence between sets. A bijection $f : X \to Y$ establishes a one-to-one correspondence between the elements of X and the elements of Y.

EXAMPLE 0.11. Define $f : \mathbb{R} \to \mathbb{R}$ by $f(x) = x^2$. Then f is not surjective since no negative numbers are in the image of f. Furthermore, f is not injective, since $1 \neq -1$, but $f(1) = f(-1)$. The image of f is the set $\mathbb{R}_+ \cup \{0\}$.

EXAMPLE 0.12. Define $f : \mathbb{R} \to \mathbb{R}$ by $f(x) = 3x + 1$. We show that this function is both injective and surjective. Hence, it is a bijection.

First, suppose that $w, x \in \mathbb{R}$ and $f(w) = f(x)$. Therefore $3w + 1 = 3x + 1$, implying that $3w = 3x$, and thus $w = x$. Since $f(w) = f(x)$ implies $w = x$ for $w, x \in \mathbb{R}$, it follows that f is injective.

Now suppose that $y \in \mathbb{R}$. Let $x = \frac{y-1}{3}$. Then

$$f(x) = 3(\tfrac{y-1}{3}) + 1 = y - 1 + 1 = y.$$

Thus for each $y \in \mathbb{R}$ there exists $x \in \mathbb{R}$ such that $f(x) = y$, implying that f is surjective.

EXAMPLE 0.13. Define $g : \mathbb{R} \to \mathbb{R}^2$ by $g(x) = (x, x)$. Then g is injective because $g(w) = g(x)$ implies that $(w, w) = (x, x)$, and therefore $w = x$. But g is not surjective because, for example, $(0, 1)$ is not in the image of g.

EXAMPLE 0.14. Define $h : \mathbb{R}^2 \to \mathbb{R}$ by $h(x_1, x_2) = x_1$. Then h is surjective because for $x \in \mathbb{R}$ we have $x = h(x, 0)$. On the other hand, h is not injective because $h(0, 0) = h(1, 0)$, but $(0, 0) \neq (1, 0)$.

To every bijective function we associate an inverse function, defined as follows:

DEFINITION 0.24. *If* $f : X \to Y$ *is a bijection, then we define a function* $f^{-1} : Y \to X$, *where* $f^{-1}(y)$ *is the unique* $x \in X$ *such that* $f(x) = y$. *The function* f^{-1} *is called the* **inverse function to** f.

EXAMPLE 0.15. In Example 0.12 the function $f : \mathbb{R} \to \mathbb{R}$, given by $f(x) = 3x + 1$, is a bijection. The inverse function to f is the function $f^{-1} : \mathbb{R} \to \mathbb{R}$ defined by

$$f^{-1}(x) = \frac{x - 1}{3}.$$

Sometimes we need to restrict a function $f : X \to Y$ to a subset A of the domain X. This is accomplished as follows:

DEFINITION 0.25. *Given $f : X \to Y$ and a subset A of X, the restriction of f to A is the function $f|_A : A \to Y$ defined by $f|_A(x) = f(x)$ for each $x \in A$.*

EXAMPLE 0.16. Consider $f : \mathbb{R} \to [-1, 1]$, defined by $f(x) = \sin(x)$, and let $A = [\frac{-\pi}{2}, \frac{\pi}{2}] \subset \mathbb{R}$. Note that f is not injective, but $f|_A$ is. In fact $f|_A$ is bijective, and its inverse function is the function given by $f|_A^{-1}(x) = \sin^{-1}(x)$.

Given a function f mapping X to Y and a function g mapping Y to Z, we can compose the two to obtain a function mapping X to Z. Specifically:

DEFINITION 0.26. *Given functions $f : X \to Y$ and $g : Y \to Z$, the composition of g with f is the function $g \circ f : X \to Z$ defined by $g \circ f(x) = g(f(x))$.*

EXAMPLE 0.17. If $f : \mathbb{R} \to \mathbb{R}$ is defined by $f(x) = 3x^2$ and $g : \mathbb{R} \to \mathbb{R}_+$ is defined by $g(x) = 5 - 2x$, then $g \circ f : \mathbb{R} \to \mathbb{R}_+$ is the function defined by

$$g \circ f(x) = 5 - 2(3x^2) = 5 - 6x^2.$$

If $f : X \to Y$ is a bijective function and $f^{-1} : Y \to X$ is its inverse, then by the definition of f^{-1} it follows that $f^{-1} \circ f(x) = x$ for all $x \in X$ and $f \circ f^{-1}(y) = y$ for all $y \in Y$.

Intuitively we think of a set X as being finite if we can count the elements of X, starting with number 1, and ending at a positive integer value. Using bijective functions, we make this idea precise as follows:

DEFINITION 0.27. *A set X is finite if X is empty or if there is a bijection $f : \{1, 2, \ldots, n\} \to X$ for some $n \in \mathbb{Z}_+$. A set that is not finite is said to be infinite.*

A special type of infinite set is the countably infinite set, a prototype of which is the set of positive integers $\mathbb{Z}_+$.

DEFINITION 0.28. *A set X is countably **infinite** if there is a bijection* $f : \mathbb{Z}_+ \to X$. *A set that is either finite or countably infinite is said to be* **countable**. *A set that is not countable is said to be* **uncountable**.

The set of integers, $\mathbb{Z}$, and the set of rational numbers, $\mathbb{Q}$, are both countable sets. The set of real numbers, $\mathbb{R}$, and the set of irrational numbers are both uncountable sets.

The following theorem provides a few important facts involving subsets, unions, and products of finite and countable sets. They will be needed at different points within the text.

THEOREM 0.29.

   (i) *A subset of a finite set is a finite set.*

   (ii) *A finite union of finite sets is a finite set.*

   (iii) *A product of finite sets is a finite set.*

   (iv) *A subset of a countable set is a countable set.*

   (v) *A countable union of countable sets is a countable set.*

   (vi) *A product of countable sets is a countable set.*

DEFINITION 0.30. *Given a set X, a **sequence in** X is a function*

$$f : \mathbb{Z}_+ \to X.$$

*The range of the function f is called the **range of the sequence**.*

Given a sequence $f : \mathbb{Z}_+ \to X$, we usually write x_n for $f(n)$ and express the sequence in the form $(x_1, x_2, ...)$ or (x_n), thinking of the sequence as a countably infinite listing of elements of X with repetition allowed in the list.

The range of a sequence can be infinite, as it is with the sequence $(1, \frac{1}{2}, \frac{1}{3}, ...)$, or it can be finite, as it is with the sequence $(1, 0, 1, 0, ...)$.

DEFINITION 0.31. *Given sequences (x_n) and (y_n), we say that (y_n) is a* **subsequence** *of (x_n) if there is a subset of the positive integers that, when put in the order $j_1 < j_2 < ... < j_n < ...$, is such that $y_n = x_{j_n}$ for all $n \in \mathbb{Z}_+$. The sequence of positive integers (j_n) is called an **indexing sequence** for the subsequence $(y_n) = (x_{j_n})$.*

EXAMPLE 0.18. The sequence $(2, 4, 6, 8, ...)$ is a subsequence of the sequence $(1, 2, 3, 4, ...)$. On the other hand, the sequence $(1, 1, 3, 3, 5, 5, ...)$ is not a subsequence of $(1, 2, 3, 4, ...)$.

Chapter 1

Topological Spaces

Topological spaces and continuous functions between them are the primary objects of study in the field of topology. In this chapter, we introduce topological spaces and some important concepts associated with them, including open sets (Section 1.1), bases (Section 1.2), and closed sets (Section 1.3). Also, in Section 1.4, we present two applications of topological spaces—one involving digital image processing, the other concerning evolutionary proximity in biology.

1.1 Open Sets and the Definition of a Topology

For many years, prior to the formalization of the field of topology, mathematicians used the concept of an open set, a simple example of which is an open interval on the real line. But over time it was realized that many of the properties held by open sets on the real line could be said to hold for certain types of subsets in any set. Eventually, the essential properties were distilled out and the concept of a collection of open sets, called a topology, evolved into the following definition:

DEFINITION 1.1. *Let X be a set. A **topology** $\mathcal{T}$ on X is a collection of subsets of X, each called an **open set**, such that*

   (i) $\varnothing$ and X are open sets;

   (ii) The intersection of finitely many open sets is an open set;

   (iii) The union of any collection of open sets is an open set.

*The set X together with a topology $\mathcal{T}$ on X is called a **topological space**.*

Thus a collection of subsets of a set X is a topology on X if it includes the empty set and X, and if finite intersections and arbitrary unions of sets in the collection are also in the collection.

IMPORTANT NOTE: *There are two things that make up a topological space: a set, X, and a collection, $\mathcal{T}$, of subsets of X that forms a topology on X. To be properly formal, we should refer to a topological space as an ordered pair $(X, \mathcal{T})$, but to simplify notation we follow the common practice of referring to the set X as a topological space, leaving it implicitly understood that there is a topology on X.*

You might have an intuition about what open sets look like if you are familiar with open intervals in the real line or open disks in the plane. Although these are open sets in what are referred to as the standard topologies on the real line and the plane, it is important to avoid typecasting open sets at this point.

Any sets at all can be an open set if we pick the topology appropriately. Let us look at a few examples.

EXAMPLE 1.1. Let X be the three-point set $\{a, b, c\}$. We consider four different collections of subsets of X in Figure 1.1 and will investigate which ones are topologies. In each case assume that the collection contains the empty set and each of the circled sets.
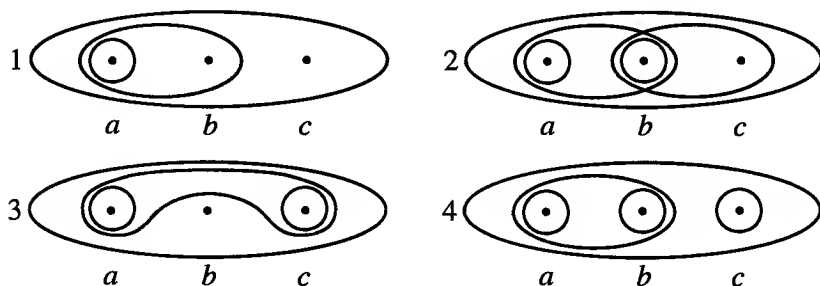
FIGURE 1.1: Which collections are topologies on $X = \{a, b, c\}$?

Both $\varnothing$ and X are in each of the four collections. We can see that for collections 1, 2, and 3, the intersection of sets in the collection is also in the collection, and the union of sets in the collection is also in the collection. Hence, collections 1, 2, and 3 depict topologies on X. However, in the case of collection 4, the sets $\{a\}$ and $\{c\}$ are each in the collection, but their union $\{a, c\}$ is not. So collection 4 does not depict a topology on X.

EXAMPLE 1.2. Let X be a nonempty set. Define $\mathcal{T} = \{\varnothing, X\}$. Notice that $\mathcal{T}$ satisfies all three of the conditions for being a topology. However, if we remove either set, we no longer have a topology. Thus $\{\varnothing, X\}$ is the minimal topology we can define on X. For obvious reasons, it is called the **trivial topology** on X.

EXAMPLE 1.3. Let X be a nonempty set and let $\mathcal{T}$ be the collection of all subsets of X. Clearly this is a topology, since unions and intersections of subsets of X are themselves subsets of X and therefore are in the collection $\mathcal{T}$. We call this the **discrete topology** on X. This is the largest topology that we can define on X.

EXAMPLE 1.4. On the real line, $\mathbb{R}$, define a topology whose open sets are the empty set and every set in $\mathbb{R}$ with a finite complement. (See Figure 1.2.) For example, $U = \mathbb{R} - \{0, 3, 7\}$ is an open set. We call this topology the **finite complement topology** on $\mathbb{R}$ and denote it by $\mathbb{R}_{fc}$.
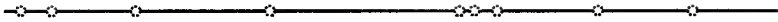
FIGURE 1.2: An open set in the finite complement topology on $\mathbb{R}$.

Let us check that $\mathbb{R}_{fc}$ is a topology:

(i) The empty set is an open set. Since the complement of $\mathbb{R}$ is empty, and therefore a finite set, it follows that $\mathbb{R}$ is an open set.

(ii) Are intersections of finitely many open sets open? Assume we have a finite collection, $U_1, \ldots, U_n$, of open sets. If any one of them is empty, then the intersection is empty and therefore is an open set. Thus assume that each U_i is nonempty. Each U_i has a finite complement, and therefore

$$U_1 = \mathbb{R} - F_1, \ U_2 = \mathbb{R} - F_2, \ldots, U_n = \mathbb{R} - F_n,$$

where each F_i is a finite set. Now, $\bigcap_{i=1}^n U_i = \mathbb{R} - \bigcup_{i=1}^n F_i$. Since a finite union of finite sets is a finite set, it follows that $\bigcap_{i=1}^n U_i$ is a subset of $\mathbb{R}$ with a finite complement and therefore is an open set. Thus, the intersection of finitely many open sets is an open set.

(iii) Are unions of open sets open? Let $\{U_\alpha\}$ be a collection of open sets. Therefore each U_α either is empty or has a finite complement. If every U_α is empty, then $\bigcup U_\alpha$ is empty, and therefore is an open set. Thus assume that at least one U_α is nonempty, say $U_{\alpha'}$. The set $U_{\alpha'}$ has a finite complement. Note that $U_{\alpha'} \subset \bigcup U_\alpha$, and it follows that $\bigcup U_\alpha$ has a finite complement since $U_{\alpha'}$ does. Therefore $\bigcup U_\alpha$ is an open set. Thus, the union of arbitrarily many open sets is an open set.

Given any nonempty set X we can similarly define the finite complement topology on X. We denote it by X_{fc}.

We have now defined three different topologies on $\mathbb{R}$: the trivial topology, the finite complement topology and the discrete topology. In the case of these three, the trivial topology, with fewest open sets, is contained within the finite complement topology, which is itself contained within the discrete topology. That is to say, every open set in the finite complement topology is also an open set in the discrete topology, but not vice versa. We say that the discrete topology is strictly finer than the finite complement topology. This terminology can be considered part of an analogy to gravel. The finer the gravel, the more ways there are to cement it together to obtain open sets. (See Figure 1.3.)

More generally, we have the following definition:

DEFINITION 1.2. *Let X be a set and let T_1 and T_2 be two topologies on X. If $T_1 \subset T_2$ then T_2 is said to be **finer** than T_1, and T_1 is said to be **coarser** than T_2. Furthermore, if T_2 is finer than T_1 but not equal to T_1 then T_2 is said to be **strictly finer** than T_1. **Strictly coarser** is defined similarly.*
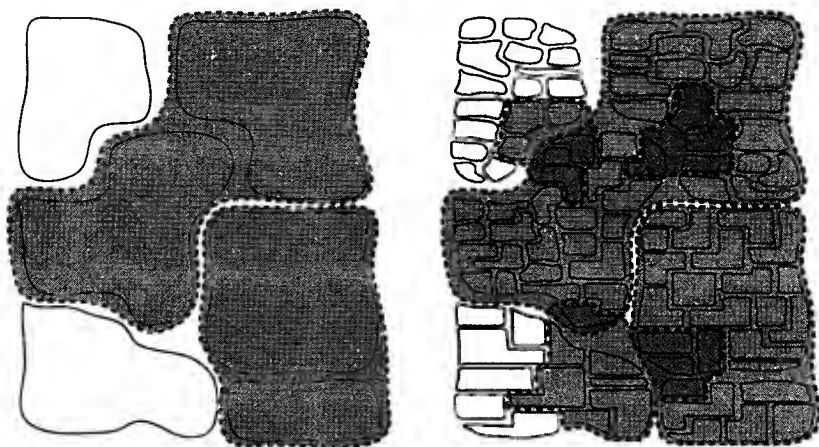
FIGURE 1.3: The finer the gravel, the greater the number of open sets that can be constructed from it.

In general, two topologies on a given set X need not be comparable. Each may contain open sets that are not open sets in the other, and therefore neither topology would be finer than the other. (See Exercise 1.5.)

To simplify discussions as we proceed, we use the following definition:

DEFINITION 1.3. *Let X be a topological space and $x \in X$. An open set U containing x is said to be a **neighborhood** of x.*

The following theorem provides us with a particularly useful means for establishing whether or not a set is open.

THEOREM 1.4. *Let X be a topological space and let A be a subset of X. Then A is open in X if and only if for each $x \in A$, there is a neighborhood U of x such that $x \in U \subset A$. (See Figure 1.4.)*
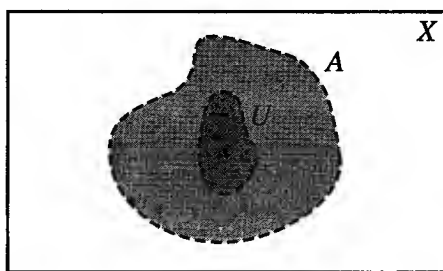
FIGURE 1.4: The set A is open in X if and only if every point in A has a neighborhood U that lies in A.

>   ***Proof.*** First suppose that A is open in X and $x \in A$. If we let $U = A$ then U is a neighborhood of x for which $x \in U \subset A$.
>
>   Now suppose that for every $x \in A$ there exists a neighborhood U_x of x such that $x \in U_x \subset A$. By the Union Lemma (Lemma 0.5), it follows that $A = \bigcup_{x \in A} U_x$. Thus, A is a union of open sets and therefore is open.   ∎

By definition, a set is open in a topological space if it is a member of the collection of sets that defines the topology. However, Theorem 1.4 provides us with an intuitive idea of what it means for a set to be open—specifically, a set is open if each point in the set has a neighborhood of points that lie in the set as well.

Exercises for Section 1.1

1.1. Determine all of the possible topologies on $X = \{a, b\}$.

1.2. On the three-point set $X = \{a, b, c\}$, the trivial topology has two open sets and the discrete topology has eight open sets. For each of $n = 3, \ldots, 7$, either find a topology on X consisting of n open sets or prove that no such topology exists.

1.3. Prove that a topology $\mathcal{T}$ on X is the discrete topology if and only if $\{x\} \in \mathcal{T}$ for all $x \in X$.

1.4. (a)   Give an example of a space where the discrete topology is the same as the finite complement topology.
   (b)   Make and prove a conjecture indicating for what class of sets the discrete and finite complement topologies coincide.

1.5. Find three topologies on the five-point set $X = \{a, b, c, d, e\}$ such that the first is finer than the second and the second finer than the third, without using either the trivial or the discrete topology. Find a topology on X that is not comparable to each of the first three that you found.

1.6. Define a topology on $\mathbb{R}$ (by listing the open sets within it) that contains the open sets $(0, 2)$ and $(1, 3)$ and that contains as few open sets as possible.

1.7. Let X be a set and assume $p \in X$. Show that the collection $\mathcal{T}$, consisting of $\varnothing$, X, and all subsets of X containing p, is a topology on X. This topology is called the **particular point topology** on X, and we denote it by PPX_p.

1.8. Let X be a set and assume $p \in X$. Show that the collection $\mathcal{T}$, consisting of $\varnothing$, X, and all subsets of X that exclude p, is a topology on X. This topology is called the **excluded point topology** on X, and we denote it by EPX_p.

1.9. Let $\mathcal{T}$ consist of $\varnothing$, $\mathbb{R}$, and all intervals $(-\infty, p)$ for $p \in \mathbb{R}$. Prove that $\mathcal{T}$ is a topology on $\mathbb{R}$.

1.2   Basis for a Topology

In all of the preceding examples of topological spaces, we were able to specify the entire collection of open sets. In general this is difficult to do, so instead we specify a smaller collection of open sets, called a basis, and then generate the rest of the open sets from this collection.

DEFINITION 1.5. *Let X be a set and $\mathcal{B}$ be a collection of subsets of X. We say $\mathcal{B}$ is a **basis** (**for a topology on** X) if the following statements hold:*

    (i) *For each x in X, there is a B in $\mathcal{B}$ such that $x \in B$.*

    (ii) *If B_1 and B_2 are in $\mathcal{B}$ and $x \in B_1 \cap B_2$, then there exists B_3 in $\mathcal{B}$ such that $x \in B_3 \subset B_1 \cap B_2$. (See Figure 1.5.)*
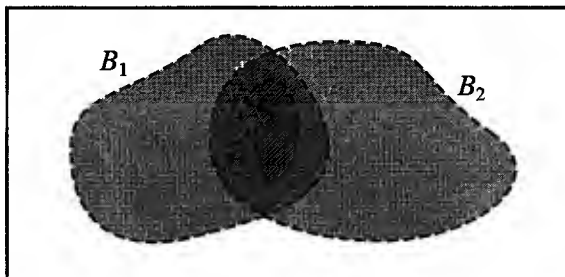
FIGURE 1.5: For every point x in the intersection of two sets in the basis, there is a set in the basis containing x, contained in the intersection.

We call the sets in $\mathcal{B}$ **basis elements.** Paraphrasing the two conditions for a basis, we have

    (i) Every point in X is contained in a basis element.

    (ii) Every point in the intersection of two basis elements is contained in a basis element contained in that intersection.

Next we describe how a basis generates a topology. First, however we look at a couple of examples of bases.

EXAMPLE 1.5. On the real line $\mathbb{R}$, let $\mathcal{B} = \{(a, b) \subset \mathbb{R} \mid a < b\}$, the set of open intervals in $\mathbb{R}$. Certainly every point of $\mathbb{R}$ is contained in an open interval and therefore is contained in a set in $\mathcal{B}$. Furthermore, if two open intervals intersect at all, they do so in an open interval, so a point in the intersection of two sets in $\mathcal{B}$ is contained in a set in $\mathcal{B}$ that is contained in the intersection. Thus $\mathcal{B}$ is a basis.

EXAMPLE 1.6. Let X be a set and let $\mathcal{B} = \{\{x\} \mid x \in X\}$. Every $x \in X$ lies in the set $\{x\}$ in $\mathcal{B}$, so the first condition for a basis is satisfied by $\mathcal{B}$. Furthermore, every pair of distinct sets in $\mathcal{B}$ is disjoint, so the second condition for a basis is satisfied automatically. Therefore $\mathcal{B}$ is a basis.

What is the advantage of having a basis on a set X? It allows us to easily define a topology.

DEFINITION 1.6. *Let $\mathcal{B}$ be a basis on a set X. The **topology** $\mathcal{T}$ **generated by** $\mathcal{B}$ is obtained by defining the open sets to be the empty set and every set that is equal to a union of basis elements.*

We need to check that the resulting collection $\mathcal{T}$ is actually a topology; we do that after looking at a couple of examples.

EXAMPLE 1.7. Let X be a nonempty set and $\mathcal{B} = \{\{x\} \mid x \in X\}$. In Example 1.6 we showed that $\mathcal{B}$ is a basis for a topology on X. What is the topology $\mathcal{T}$ that $\mathcal{B}$ generates? Notice that every subset U of X is the union of the single-point subsets corresponding to its elements. Therefore every subset of X is an open set in $\mathcal{T}$, implying that $\mathcal{B}$ generates the discrete topology on X.

EXAMPLE 1.8. On the real line $\mathbb{R}$, let $\mathcal{B} = \{(a, b) \subset \mathbb{R} \mid a < b\}$. In Example 1.5 we showed that $\mathcal{B}$ is a basis for a topology on $\mathbb{R}$. The topology generated by $\mathcal{B}$ is called the **standard topology** on $\mathbb{R}$ and is the topology most commonly used on the real line. Open sets in the standard topology on $\mathbb{R}$ are unions of open intervals. When we refer to $\mathbb{R}$ as a topological space, unless otherwise specified, it will always be assumed that the topology is the standard topology.

Next we show that the topology generated by a basis actually is a topology. To begin we need the following lemma.

LEMMA 1.7. *Let $\mathcal{B}$ be a basis. Assume that $B_1, \ldots, B_n \in \mathcal{B}$ and that $x \in \bigcap_{i=1}^{n} B_i$. Then there exists $B' \in \mathcal{B}$ such that $x \in B' \subset \bigcap_{i=1}^{n} B_i$.*

> ***Proof.*** We prove this by induction on n, starting with $n = 2$. The $n = 2$ case holds by the second condition in the definition of a basis.
>
> Assume that the result is true for $n - 1$. Suppose that the sets $B_1, \ldots, B_n$ are in $\mathcal{B}$ and that $x \in \bigcap_{i=1}^{n} B_i$. Then $x \in \bigcap_{i=1}^{n-1} B_i$, and the induction hypothesis implies that there exists $B^* \in \mathcal{B}$ such that $x \in B^* \subset \bigcap_{i=1}^{n-1} B_i$. Now $x \in B^* \cap B_n$; therefore by the second condition in the definition of a basis, there exists $B' \in \mathcal{B}$ such that $x \in B' \subset B^* \cap B_n$. Since $B^* \subset \bigcap_{i=1}^{n-1} B_i$ it follows that $x \in B' \subset \bigcap_{i=1}^{n} B_i$. Thus, if the result holds for $n - 1$, then it holds for n, and by induction the lemma follows. ∎

THEOREM 1.8. *The topology $\mathcal{T}$ generated by a basis $\mathcal{B}$ is a topology.*

> ***Proof.*** The empty set $\varnothing$ is in $\mathcal{T}$ by definition. Since every point in X is contained in some basis element, it follows that X is the union of all of the basis elements and therefore is in $\mathcal{T}$.
>
> Next we show that a finite intersection of sets in $\mathcal{T}$ is in $\mathcal{T}$. Thus, let $V = U_1 \cap \ldots \cap U_n$ where each U_i is in $\mathcal{T}$. If any one of the U_i is empty, then so is V, and in this case V is in $\mathcal{T}$. Thus assume that each U_i is a union of basis elements. We show that V is a union of basis elements

as well. Let $x \in V$ be arbitrary. Then $x \in U_i$ for all i. Since each U_i is a union of basis elements, there exists a basis element B_i such that $x \in B_i \subset U_i$ for each i. Then $x \in \bigcap_{i=1}^{n} B_i$; therefore, by Lemma 1.7, there exists a basis element B_x such that $x \in B_x \subset \bigcap_{i=1}^{n} B_i \subset V$. It follows from the Union Lemma (Lemma 0.5) that $V = \bigcup_{x \in V} B_x$, and therefore V is a union of basis elements. Thus, a finite intersection of sets in $\mathcal{T}$ is in $\mathcal{T}$.

Finally, we show that an arbitrary union of sets in $\mathcal{T}$ is in $\mathcal{T}$. Let $V = \bigcup U_\alpha$ where each U_α is either the empty set or a union of basis elements. If each U_α is empty, then so is V; on the other hand, if at least one U_α is nonempty, then V is a union of basis elements, since it is the union of all of the basis elements making up the U_α's. Therefore an arbitrary union of sets in $\mathcal{T}$ is in $\mathcal{T}$.

Thus, the collection of sets $\mathcal{T}$ is a topology, and we are justified in calling it the topology generated by the basis $\mathcal{B}$. ∎

IMPORTANT NOTE: *Not only does a basis generate a topology, but each basis element is also itself an open set in the topology generated by the basis.*

EXAMPLE 1.9. On $\mathbb{R}$, let $\mathcal{B} = \{[a, b) \subset \mathbb{R} \mid a < b\}$. The collection $\mathcal{B}$ is a basis for a topology on $\mathbb{R}$. (See Exercise 1.10.) We call the topology generated by this basis the **lower limit topology** since each basis element contains its lower limit. We denote $\mathbb{R}$ with this topology by $\mathbb{R}_l$.

The intervals $[0, 2)$ and $(0, 2)$ are both open in $\mathbb{R}_l$. The former is open since it is a basis element; the latter is open since it is the union of the basis elements $B_i = [\frac{1}{i}, 2)$, where $i = 1, 2, 3, \ldots$.

We can similarly define the **upper limit topology** on $\mathbb{R}$ via the basis $\mathcal{B} = \{(a, b] \subset \mathbb{R} \mid a < b\}$.

We have now introduced six topologies on the real line: the standard topology, the upper limit topology, the lower limit topology, the finite complement topology, the discrete topology, and the trivial topology. Of course, the trivial topology is the coarsest of these topologies, and the discrete topology is the finest, but how do the rest of these topologies compare? We ask you to investigate this question in Exercise 1.13.

On the set of integers, $\mathbb{Z}$, the standard topology turns out to be the discrete topology. (We will see why in Section 3.1.) There are, however, nondiscrete topologies on $\mathbb{Z}$ that play a role in topology and its applications. In the next example we introduce a topology on $\mathbb{Z}$ that is useful in digital image processing. We discuss this application further in Sections 1.4 and 11.3.

EXAMPLE 1.10. For each $n \in \mathbb{Z}$, define

$$B(n) = \begin{cases} \{n\} & \text{if } n \text{ is odd,} \\ \{n - 1, n, n + 1\} & \text{if } n \text{ is even.} \end{cases}$$

We illustrate these sets in Figure 1.6. The collection $\mathcal{B} = \{B(n) \mid n \in \mathbb{Z}\}$ is a basis for a topology on $\mathbb{Z}$. (See Exercise 1.14.) The resulting topology is called

the **digital line topology**, and we refer to $\mathbb{Z}$ with this topology as the **digital line**.
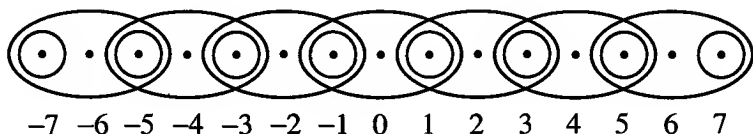
FIGURE 1.6: Basis elements for the digital line topology.

By definition, the open sets in the topology generated by a basis are the sets obtained by taking unions of basis elements. The next theorem provides another simple way to describe the sets in such a topology.

THEOREM 1.9. *Let X be a set and $\mathcal{B}$ be a basis for a topology on X. Then U is open in the topology generated by $\mathcal{B}$ if and only if for each $x \in U$ there exists a basis element $B_x \in \mathcal{B}$ such that $x \in B_x \subset U$. (See Figure 1.7.)*
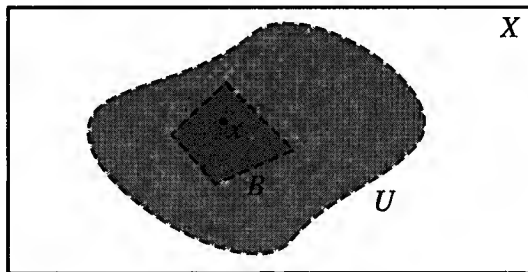
FIGURE 1.7: The set U is open if and only if every x in U is contained in a basis element contained in U.

Proof. Suppose that U is an open set in the topology generated by $\mathcal{B}$ and that $x \in U$. Since U is a union of basis elements, there is at least one basis element B' making up this union that contains x. Clearly then $x \in B' \subset U$.

Now suppose that $U \subset X$ is such that for each $x \in U$ there exists $B_x \in \mathcal{B}$ such that $x \in B_x \subset U$. By the Union Lemma (Lemma 0.5), $U = \bigcup_{x \in U} B_x$, and therefore U is a union of basis elements. Thus, U is an open set in the topology generated by $\mathcal{B}$. ∎

EXAMPLE 1.11. We have examined a few topologies on the real line. Next, let us look at the plane $\mathbb{R}^2$. For $p = (p_1, p_2)$ and $q = (q_1, q_2)$, two points in $\mathbb{R}^2$, we introduced the Euclidean distance formula

$$d(p, q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2}$$

in Section 0.4. For each x in $\mathbb{R}^2$ and $\varepsilon > 0$, define

$$B(x, \varepsilon) = \{p \in \mathbb{R}^2 \mid d(x, p) < \varepsilon\}.$$

The set $B(x, \varepsilon)$ is called the **open ball of radius ε centered at x**. Let

$$\mathcal{B} = \{B(x, \varepsilon) \mid x \in \mathbb{R}^2, \varepsilon > 0\}.$$

So $\mathcal{B}$ is the collection of all open balls associated with the Euclidean distance d. In Theorem 1.10 we show that $\mathcal{B}$ is a basis for a topology on $\mathbb{R}^2$. We call the topology generated by $\mathcal{B}$ the **standard topology on $\mathbb{R}^2$**. It is the most common topology used on $\mathbb{R}^2$.

THEOREM 1.10. *The collection $\mathcal{B} = \{B(x, \varepsilon) \mid x \in \mathbb{R}^2, \varepsilon > 0\}$ is a basis for a topology on $\mathbb{R}^2$.*

Before proceeding with the proof of this theorem, we prove the following supporting lemma:

LEMMA 1.11. *Let y be in $\mathbb{R}^2$ and assume $r > 0$. Then for every $x \in B(y, r)$ there exists an $\varepsilon > 0$ such that $B(x, \varepsilon) \subset B(y, r)$.*

The lemma indicates that if a point x is in some open ball B in $\mathbb{R}^2$, then there is an open ball centered at x contained in B as well. (See Figure 1.8.) We will use the lemma in establishing that the collection of open balls in $\mathbb{R}^2$ satisfies the second property of a basis.
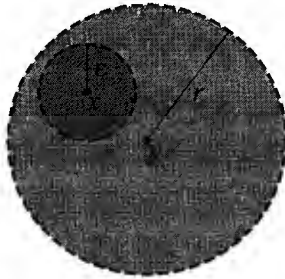
FIGURE 1.8: Every point x in $B(y, r)$ is the center of some open ball contained in $B(y, r)$.

Proof of Lemma 1.11. Suppose $x \in B(y, r)$. Then $d(x, y) < r$. Choose ε such that $0 < \varepsilon < r - d(x, y)$. We claim that $B(x, \varepsilon) \subset B(y, r)$. Suppose $z \in B(x, \varepsilon)$. Then we have

$$\begin{aligned} d(y, z) &< d(y, x) + d(x, z) \\ &< d(y, x) + \varepsilon \\ &< d(y, x) + r - d(x, y) \\ &= r. \end{aligned}$$

Therefore $z \in B(y, r)$, and it follows that $B(x, \varepsilon) \subset B(y, r)$, completing the proof of the lemma. ∎

Proof of Theorem 1.10. Since each $x \in \mathbb{R}^2$ is contained in the basis element $B(x, 1)$, the first condition for a basis is satisfied.

Now we need to verify that, if x is in the intersection of two basis elements, there is then a basis element containing x and contained in the intersection. Suppose $x \in B(p, r_1) \cap B(q, r_2)$. By Lemma 1.11 there exist $\varepsilon_1, \varepsilon_2 > 0$ such that $B(x, \varepsilon_1) \subset B(p, r_1)$ and $B(x, \varepsilon_2) \subset B(q, r_2)$. Let $\varepsilon = \min\{\varepsilon_1, \varepsilon_2\}$. Then

$$B(x, \varepsilon) \subset B(x, \varepsilon_1) \cap B(x, \varepsilon_2) \subset B(p, r_1) \cap B(q, r_2),$$

and it follows that $\mathcal{B}$ satisfies the second condition for a basis.

Therefore $\mathcal{B}$ is a basis for a topology on $\mathbb{R}^2$.   ■

In the topology on $\mathbb{R}^2$ generated by $\mathcal{B}$ the open sets are sets that can be expressed as a union of open balls. For example, open balls, open rectangles, and open half planes, as illustrated in Figure 1.9, are all open sets. (See Exercise 1.17.)
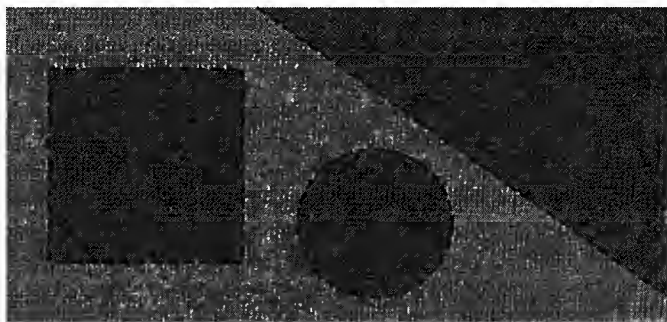
FIGURE 1.9: Open sets in the standard topology on $\mathbb{R}^2$.

We can similarly define the **standard topology on $\mathbb{R}^n$** using the Euclidean distance formula

$$d(p, q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \ldots + (p_n - q_n)^2}.$$

We define an open ball to be a set of the form

$$B(x, \varepsilon) = \{p \in \mathbb{R}^n \mid d(x, p) < \varepsilon\},$$

and we define a basis $\mathcal{B}$ for the topology to be the collection of all open balls in $\mathbb{R}^n$ given by $\mathcal{B} = \{B(x, \varepsilon) \mid x \in \mathbb{R}^n, \ \varepsilon > 0\}$.

Not only is the standard topology on $\mathbb{R}^2$ generated by the collection of open balls in $\mathbb{R}^2$, but, as the following theorem indicates, it is also generated by the collection of open rectangles in $\mathbb{R}^2$, where an **open rectangle** is a set of the form $(a, b) \times (c, d)$.

THEOREM 1.12. *On the plane $\mathbb{R}^2$, let*

$$\mathcal{B} = \{(a, b) \times (c, d) \subset \mathbb{R}^2 \mid a < b, \ c < d\}.$$

Then $\mathcal{B}$ is a basis, and the topology $\mathcal{T}'$ generated by $\mathcal{B}$ is the standard topology on $\mathbb{R}^2$.

Proof. See Exercise 1.16.   ■

Theorem 1.12 demonstrates that a given topology can be generated by more than one basis. Furthermore, it demonstrates that the geometric shape of the individual basis elements is not significant in determining the open sets in the topology. To be specific, the topology on the plane generated by the collection of open balls is the same as that generated by the collection of open rectangles. We can also generate that same topology by a collection of diamonds or even hearts, as illustrated in Figure 1.10.

FIGURE 1.10: Many different bases can generate the same topology.

In a fashion similar to that used in Theorem 1.12, it is not difficult to show that the standard topology on $\mathbb{R}^n$ is also generated by a basis consisting of open rectangles $(a_1, b_1) \times \ldots \times (a_n, b_n)$.

The following theorem provides us with a quick way of proving that a collection of open sets in a topology is a basis for the topology. We will find this theorem useful when we introduce some new constructions for topologies in Chapter 3.

THEOREM 1.13. *Let X be a set with topology $\mathcal{T}$, and let $\mathcal{C}$ be a collection of open sets in X. If, for each open set U in X and for each $x \in U$, there is an open set V in $\mathcal{C}$ such that $x \in V \subset U$, then $\mathcal{C}$ is a basis that generates the topology $\mathcal{T}$.*

> **Proof.** First, we check that $\mathcal{C}$ is a basis. Let x be in X. Since X is itself an open set, there is an open set V in $\mathcal{C}$ such that $x \in V \subset X$. Therefore every point in X is contained in some open set V in the collection $\mathcal{C}$. Thus the first condition for a basis is satisfied.
>
> Suppose now that x is in the intersection of two of the open sets in $\mathcal{C}$, V_1 and V_2. Then $x \in V_1 \cap V_2$, which is itself an open set. According to our hypothesis, there must be an open set V_3 in $\mathcal{C}$ such that $x \in V_3 \subset V_1 \cap V_2$. Therefore $\mathcal{C}$ is a basis.
>
> Now, we need to check that $\mathcal{T}'$, the topology generated by $\mathcal{C}$, coincides with $\mathcal{T}$. Suppose U is open in $\mathcal{T}$. Then by the hypothesis, for every $x \in U$ there is an open set V in $\mathcal{C}$ such that $x \in V \subset U$. So by Theorem 1.9, U is open in $\mathcal{T}'$, the topology generated by $\mathcal{C}$. Therefore $\mathcal{T} \subset \mathcal{T}'$. Now suppose that W is an open set in $\mathcal{T}'$. Then W is a union of open sets in $\mathcal{C}$, all of which are open in $\mathcal{T}$. But every union of open sets in $\mathcal{T}$ is open in $\mathcal{T}$, so W is also open in $\mathcal{T}$. Thus $\mathcal{T}' \subset \mathcal{T}$ and it follows that $\mathcal{T}$ and $\mathcal{T}'$ coincide.   ∎

EXAMPLE 1.12. We can immediately see from Theorem 1.13 that the open hearts in Figure 1.10 form a basis for the standard topology on $\mathbb{R}^2$. First, note that the hearts themselves are open sets in the standard topology. Now, suppose that x is in an open set U. Then there is an open ball B containing x, contained in U. If we choose a heart H containing x that is small enough to fall entirely in B, then we have $x \in H \subset U$. Therefore, by Theorem 1.13, the collection of hearts is a basis for the standard topology on $\mathbb{R}^2$.

Exercises for Section 1.2

1.10. Show that $\mathcal{B} = \{[a, b) \subset \mathbb{R} \mid a < b\}$ is a basis for a topology on $\mathbb{R}$.

1.11. Determine which of the following collections of subsets of $\mathbb{R}$ are bases:
   (a) $\mathcal{C}_1 = \{(n, n + 2) \subset \mathbb{R} \mid n \in \mathbb{Z}\}$
   (b) $\mathcal{C}_2 = \{[a, b] \subset \mathbb{R} \mid a < b\}$
   (c) $\mathcal{C}_3 = \{[a, b] \subset \mathbb{R} \mid a \leq b\}$
   (d) $\mathcal{C}_4 = \{(-x, x) \subset \mathbb{R} \mid x \in \mathbb{R}\}$
   (e) $\mathcal{C}_5 = \{(a, b) \cup \{b + 1\} \subset \mathbb{R} \mid a < b\}$

1.12. Determine which of the following are open sets in $\mathbb{R}_l$. In each case, prove your assertion.
$$A = [4, 5) \quad B = \{3\} \quad C = [1, 2] \quad D = (7, 8)$$

1.13. Consider the following six topologies defined on $\mathbb{R}$: the trivial topology, the discrete topology, the finite complement topology, the standard topology, the lower limit topology, and the upper limit topology. Show how they compare to each other (finer, strictly finer, coarser, strictly coarser, noncomparable) and justify your claim.

1.14. Let $\mathcal{B}$ be the collection of subsets of $\mathbb{Z}$ used in defining the digital line in Example 1.10. Show that $\mathcal{B}$ is a basis for a topology on $\mathbb{Z}$.

1.15. An **arithmetic progression** in $\mathbb{Z}$ is a set
$$A_{a,b} = \{\ldots, a - 2b, a - b, a, a + b, a + 2b, \ldots\}$$

with $a, b \in \mathbb{Z}$ and $b \neq 0$. Prove that the collection of arithmetic progressions

$$\mathcal{A} = \{A_{a,b} \mid a, b \in \mathbb{Z} \text{ and } b \neq 0\}$$

is a basis for a topology on $\mathbb{Z}$. The resulting topology is called the **arithmetic progression topology** on Z.

1.16. Prove Theorem 1.12: On the plane $\mathbb{R}^2$, let

$$\mathcal{B} = \{(a, b) \times (c, d) \subset \mathbb{R}^2 \mid a < b, \ c < d\}.$$

   (a) Show that $\mathcal{B}$ is a basis for a topology on $\mathbb{R}^2$.
   (b) Show that the topology, $\mathcal{T}'$, generated by $\mathcal{B}$ is the standard topology on $\mathbb{R}^2$. (Hint: if $\mathcal{T}$ is the standard topology, show that $\mathcal{T} \subset \mathcal{T}'$ and $\mathcal{T}' \subset \mathcal{T}$.)

1.17. An **open half plane** is a subset of $\mathbb{R}^2$ in the form $\{(x, y) \in \mathbb{R}^2 \mid Ax + By < C\}$ for some $A, B, C \in \mathbb{R}$ with either A or B nonzero. (See Figure 1.11.) Prove that open half planes are open sets in the standard topology on $\mathbb{R}^2$.
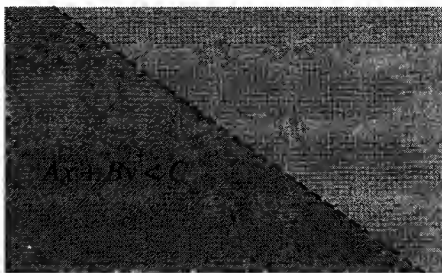
FIGURE 1.11: An open half plane in $\mathbb{R}^2$.

1.18. Show that the collection $\{(-\infty, q) \subset \mathbb{R} \mid q \text{ rational}\}$ is a basis for the topology in Exercise 1.9.

1.19. (a) Show that the collection $\{\{a\} \times (b, c) \subset \mathbb{R}^2 \mid a, b, c \in \mathbb{R}\}$ of vertical intervals in the plane is a basis for a topology on $\mathbb{R}^2$. We call this topology the **vertical interval topology**.

    (b) Compare the vertical interval topology with the standard topology on $\mathbb{R}^2$.

Supplementary Exercises: Creating a Topology from a Collection of Subsets

Throughout these exercises, let X be a set and let $\mathcal{C}$ be a collection of subsets of X whose union equals X. Of course, in general $\mathcal{C}$ is not a topology. However, we would like to be able to construct a topology so that the subsets in $\mathcal{C}$ are open sets in the topology. We could just take the discrete topology, where all subsets of X are open. However, we prefer to construct a topology containing $\mathcal{C}$ with as few extra open sets as possible, a "smallest" topology containing $\mathcal{C}$.

Let $\mathcal{B}_\mathcal{C}$ be the collection of all subsets of X that can be expressed as an intersection of finitely many of the sets from $\mathcal{C}$.

SE 1.20. Show that $\mathcal{B}_\mathcal{C}$ is a basis for a topology on X.

Let $\mathcal{T}_\mathcal{C}$ be the topology generated by the basis $\mathcal{B}_\mathcal{C}$. The topology $\mathcal{T}_\mathcal{C}$ consists of arbitrary unions of finite intersections of sets in $\mathcal{C}$. The collection $\mathcal{C}$ is what is known as a **subbasis** for the topology $\mathcal{T}_\mathcal{C}$.

SE 1.21. Prove that every set in $\mathcal{C}$ is an open set in the topology $\mathcal{T}_\mathcal{C}$; that is, prove that $\mathcal{C} \subset \mathcal{T}_\mathcal{C}$.

Now we verify that $\mathcal{T}_\mathcal{C}$ is the smallest topology containing $\mathcal{C}$.

SE 1.22. Let $\mathcal{T}$ be a topology on X containing $\mathcal{C}$. Thus assume that every set in $\mathcal{C}$ is an open set in $\mathcal{T}$. Prove that $\mathcal{T}_\mathcal{C} \subset \mathcal{T}$.

For the remaining exercises assume that each $x \in X$ is contained in at most finitely many sets in $\mathcal{C}$. (For example, this occurs if X is a finite set.) For each $x \in X$, let B_x be the intersection of the sets in $\mathcal{C}$ containing x.

SE 1.23. Show that the collection $\mathcal{B}'_\mathcal{C} = \{B_x\}_{x \in X}$ is a basis for $\mathcal{T}_\mathcal{C}$.

In general, a topological space does not have a minimal basis, that is, a basis containing as few basis elements as possible. However, in the setting we are considering here, where each $x \in X$ is contained in finitely many of the sets in $\mathcal{C}$, we can show $\mathcal{B}'_\mathcal{C}$ is the minimal basis for $\mathcal{T}_\mathcal{C}$ as follows:

SE 1.24. Prove that if $\mathcal{B}$ is a basis for $\mathcal{T}_\mathcal{C}$, then $\mathcal{B}'_\mathcal{C} \subset \mathcal{B}$.

1.3   Closed Sets

So far, we have been talking exclusively about open sets. Now it is time to look at the complementary concept and introduce closed sets. Their definition is quite simple.

DEFINITION 1.14. *A subset A of a topological space X is **closed** if the set $X - A$ is open.*

EXAMPLE 1.13. Consider $\mathbb{R}$ with the standard topology. Then, as illustrated in Figure 1.12, since $(0, 1)$ is open, $(-\infty, 0] \cup [1, \infty)$ is closed; since $(-\infty, a) \cup (b, \infty)$ is open, $[a, b]$ is closed; and, since $(-\infty, c) \cup (c, \infty)$ is open, $\{c\}$ is closed.
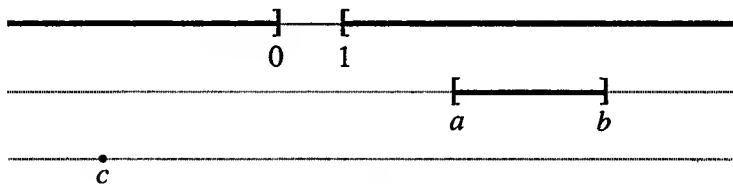
FIGURE 1.12: Closed sets in $\mathbb{R}$ with the standard topology.

DEFINITION 1.15.

(i) *For each x in $\mathbb{R}^2$ and $\varepsilon > 0$, define the **closed ball of radius ε centered at x** to be the set*

$$\bar{B}(x, \varepsilon) = \{y \in \mathbb{R}^2 \mid d(x, y) \leq \varepsilon\},$$

where $d(x, y)$ is the Euclidean distance between x and y.

(ii) *If $[a, b]$ and $[c, d]$ are closed bounded intervals in $\mathbb{R}$, then the product $[a, b] \times [c, d] \subset \mathbb{R}^2$ is called a **closed rectangle**.*

THEOREM 1.16. *Closed balls and closed rectangles are closed sets in the standard topology on $\mathbb{R}^2$.*

Proof. We ask you to prove that closed balls are closed sets in Exercise 1.26. Here, we prove that closed rectangles are closed sets in the standard topology on $\mathbb{R}^2$.

Let $A = [a, b] \times [c, d]$ be a closed rectangle in $\mathbb{R}^2$. To show that A is a closed set in the standard topology, we prove that $\mathbb{R}^2 - A$ is an open set. Note that $\mathbb{R}^2 - A$ can be expressed as the union of four open half-planes: $\{(x, y) \mid x < a\}$, $\{(x, y) \mid x > b\}$, $\{(x, y) \mid y < c\}$, and $\{(x, y) \mid y > d\}$. Since each of these half-planes is an open set (see Exercise 1.17), and a union of open sets is an open set, it follows that $\mathbb{R}^2 - A$ is an open set. Hence, the rectangle A is a closed set. ∎

EXAMPLE 1.14. Consider $\mathbb{R}^2$ with the standard topology. In Figure 1.13 the top three sets are closed, and the bottom three are not.
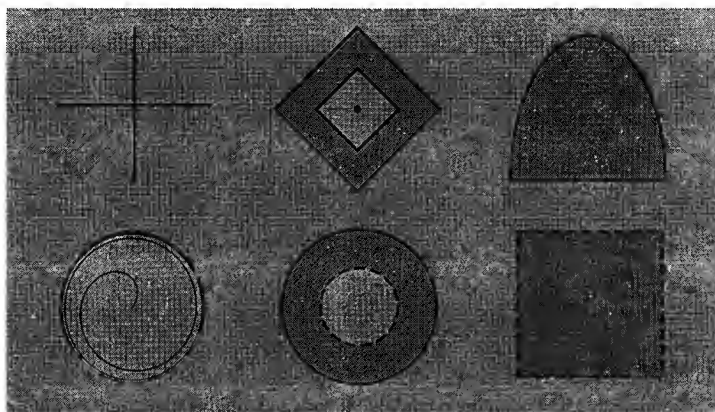
FIGURE 1.13: In $\mathbb{R}^2$ with the standard topology, the top three sets are closed, but the bottom three are not.

The spiral on the lower left is the graph of the polar-coordinates equation $r = \frac{\theta}{\theta+1}$ with $\theta \geq 0$. It winds out toward the circle S^1, but it contains no point from S^1. The spiral is not closed because the point $(1, 0)$ is in the complement, but no open ball containing $(1, 0)$ is a subset of the complement since every such open ball intersects the spiral. Thus the complement is not open, implying that the spiral is not closed.

If a set C is closed, then by definition its complement is open. What can we say about the complement of an open set? Let U be an open set, and let $K = X - U$ be its complement. Consider the complement of K. We have $X - K = X - (X - U) = U$, and U is an open set. Thus K, the complement of U, is closed. Therefore the complement of an open set is closed.

In summary, the complement of a closed set is open by the definition of closed set, and the complement of an open set is closed by the argument just presented. (See Figure 1.14.)

EXAMPLE 1.15. Let X be a set with the discrete topology. Every subset of X is an open set. Given an arbitrary set A in X, its complement is then open

since every set is open. Therefore A is closed and, since A was arbitrary, it follows that every subset of X is also a closed set in the discrete topology.
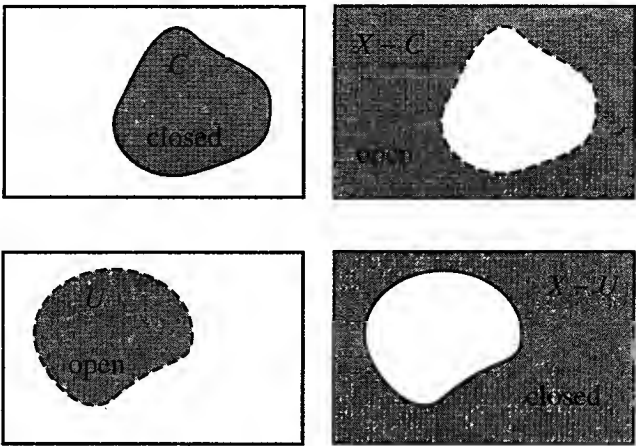
FIGURE 1.14: If C is closed, its complement is open; if U is open, its complement is closed.

EXAMPLE 1.16. Let the set $\{a, b, c, d\}$ have the topology $\mathcal{T}$ as shown in Figure 1.15. Note that $\{b\}$ is open and not closed, $\{a\}$ is closed and not open, $\{a, b\}$ is both open and closed, and $\{b, c\}$ is neither open nor closed.
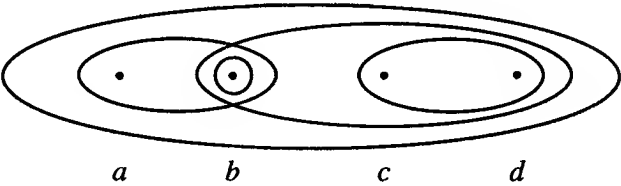
FIGURE 1.15: A topology with open sets, closed sets, sets both open and closed, and sets neither open nor closed.

IMPORTANT NOTE: *A common error is to mistake* closed *for* not open, *but as this last example illustrates, sets can be any combination of open and closed.*

Here is a riddle to help remember this:

Question: How is a subset different from a door?

Answer: A door must be open or closed. But a subset can be either, both, or neither.

It is hard to resist math humor, but in the future, we will try.

Since closed sets are complementary to open sets, their properties are similar, but there are some fundamental differences.

THEOREM 1.17. *Let X be a topological space. The following statements about the collection of closed sets in X hold:*

   (i) $\varnothing$ *and X are closed.*

   (ii) *The intersection of any collection of closed sets is a closed set.*

   (iii) *The union of finitely many closed sets is a closed set.*

   Proof. See Exercise 1.33. ∎

Thus, the collection of closed sets in a topological space X includes the empty set and X and is such that finite unions and arbitrary intersections of closed sets are closed sets.

All single-point sets are closed in $\mathbb{R}^n$ with the standard topology. However, there are topologies where this is not the case. For example, consider the topology $\mathcal{T}^*$ on $\mathbb{R}$ generated by the basis of sets $\{[n, n+1) \subset \mathbb{R} \mid n \in \mathbb{Z}\}$. It is not difficult to see that no single-point set is closed in $\mathcal{T}^*$.

Next we present a property of topological spaces that has some particularly nice consequences for spaces satisfying it. As we show below, these consequences include having single-point sets closed. We will see a few of the other convenient consequences of this property throughout the text.

DEFINITION 1.18. *A topological space X is **Hausdorff** if for every pair of distinct points x and y in X, there exist disjoint neighborhoods U and V of x and y, respectively. (See Figure 1.16.)*
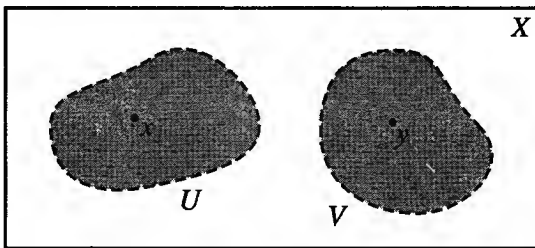
FIGURE 1.16: In a Hausdorff space distinct points have disjoint neighborhoods.

In his 1914 book, *Grundzüge der Mengenlehre*, Felix Hausdorff presented the first general axiom system for what was to become known as a topological space. His axioms essentially describe what we refer to as a basis for a topology. His system also included an axiom that is no longer contained in the axioms for a general topological space, but that now defines the special class of topological spaces that are known—appropriately enough—as Hausdorff spaces.

The Hausdorff property is easy to remember: In a Hausdorff space, each point in a pair of points can be "housed off" from the other by disjoint neighborhoods.

EXAMPLE 1.17. The real line $\mathbb{R}$ with the standard topology is Hausdorff. Given two distinct points a and b, there are disjoint open intervals containing them. For example, if $a < b$, then the intervals

$$U = \left(a - 1, \frac{(a+b)}{2}\right) \text{ and } V = \left(\frac{(a+b)}{2}, b+1\right)$$

are disjoint and contain a and b, respectively.

EXAMPLE 1.18. Every set X with the discrete topology is Hausdorff. If x and y are distinct points, then the sets $\{x\}$ and $\{y\}$ are disjoint neighborhoods of x and y, respectively.

As we have already indicated, Hausdorff spaces have a variety of useful properties, such as:

THEOREM 1.19. *If X is a Hausdorff space, then every single-point subset of X is closed.*

> **Proof.** Suppose $x \in X$. In order to show that $\{x\}$ is closed, we show that $X - \{x\}$ is open. Let $y \in X - \{x\}$ be arbitrary. Since X is Hausdorff, there are disjoint neighborhoods U and V containing x and y, respectively. It follows that $x \notin V$ and therefore $y \in V \subset X - \{x\}$. Since every $y \in X - \{x\}$ is in an open set contained in $X - \{x\}$, Theorem 1.4 implies that $X - \{x\}$ is open. ∎

Exercises for Section 1.3

1.25. Prove that, in a topological space X, if U is open and C is closed, then $U - C$ is open and $C - U$ is closed.

1.26. Prove that closed balls are closed sets in the standard topology on $\mathbb{R}^2$.

1.27. The **infinite comb** C is the subset of the plane illustrated in Figure 1.17 and defined by

$$C = \{(x, 0) \mid 0 \leq x \leq 1\} \cup \{(\tfrac{1}{2^n}, y) \mid n = 0, 1, 2, \ldots \text{ and } 0 \leq y \leq 1\}.$$

(a) Prove that C is not closed in the standard topology on $\mathbb{R}^2$.
(b) Prove that C is closed in the vertical interval topology on $\mathbb{R}^2$. (See Exercise 1.19.)
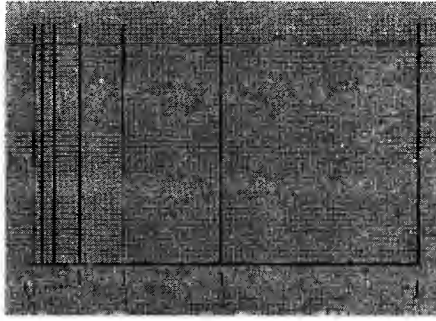
FIGURE 1.17: The infinite comb.

1.28. Which sets are closed sets in the finite complement topology on a topological space X?

1.29. Which sets are closed sets in the excluded point topology EPX_p on a set X? (See Exercise 1.8.)

1.30. Which sets are closed sets in the particular point topology PPX_p on a set X? (See Exercise 1.7.)

1.31. Show that a single-point set $\{n\}$ is closed in the digital line topology if and only if n is even.

1.32. Prove that intervals of the form $[a, b)$ are closed in the lower limit topology on $\mathbb{R}$.

1.33. Prove Theorem 1.17: Let X be a topological space.
    (a) Prove that $\varnothing$ and X are closed sets.
    (b) Prove that the intersection of any collection of closed sets in X is a closed set.
    (c) Prove that the union of finitely many closed sets in X is a closed set.

1.34. On the five-point set $\{a, b, c, d, e\}$ construct two topologies, one that is Hausdorff (other than the discrete topology) and one that is not Hausdorff (other than the trivial topology).

1.35. Show that $\mathbb{R}$ in the lower limit topology is Hausdorff.

1.36. Show that $\mathbb{R}$ in the finite complement topology is not Hausdorff.

1.37. Prove that the arithmetic progression topology on $\mathbb{Z}$, introduced in Exercise 1.15, is Hausdorff.

1.4   *Examples of Topologies in Applications*

In this section we discuss two applications of topological spaces: one showing how topological spaces model digital image displays in digital topology, and the other showing how topological spaces are employed to model evolutionary proximity in molecular biology.

Digital Topology

Digital topology is the study of topological relationships on a digital image display (for example, a computer screen). It plays a role in the field of digital image processing. The digital image display contains a rectangular pixel array, as illustrated in Figure 1.18. In digital topology, this array is modeled by what is known as the digital plane. We provide a specific definition hereafter.
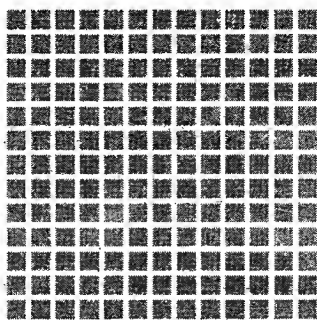
FIGURE 1.18: The digital image display is a rectangular array of pixels.

An important task in digital image processing is to determine features of an object from a digital image of it. For example, an optical character-recognition program reads a digital image of a character (such as in Figure 1.19) and attempts to determine the character represented by the image so that subsequently it can be used within a word-processing program. Topological properties of the image assist in determining the intended character. The fact that the image in Figure 1.19 encloses two regions helps distinguish the associated character from characters that do not.
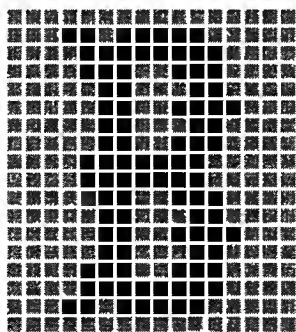
FIGURE 1.19: A digital image of the letter B.

In the initial development of the field of digital topology, work focused on defining and studying digital analogs of topological concepts (for example, connectedness and continuity) without having an underlying topology on the model of the digital image display. The 1979 paper, "Digital Topology," by Azriel Rosenfeld (1931–2004), was one of the first introductions to these ideas. Later, topological spaces were found that appropriately model the digital image

display, enabling topological concepts to be used directly in the digital domain. (See [Kon], for example.)

In this section we introduce the digital line and digital plane models and consider some basic topological properties associated with them. In Section 11.3 we examine an application of digital topology to digital image processing.

The digital line, introduced in Example 1.10, models a one-dimensional digital image display. We assume that we have an infinite line of pixels, each of which corresponds to an odd integer. For each odd integer we have a basis element $B(n) = \{n\}$. Consequently, each individual pixel is an open set in the digital line topology. In the digital line we also have structure that reflects the adjacency shared by consecutive pixels. We regard each even integer n as representing the boundary between the pixels at $n - 1$ and $n + 1$. Associated to each even n we have a basis element $B(n) = \{n - 1, n, n + 1\}$.

In the digital line, every odd integer is an open set, and every even integer is a closed set (see Exercise 1.31). We view the digital line as a set of open pixels corresponding to the odd integers, along with the set of closed boundaries between the pixels corresponding to the even integers. (See Figure 1.20.)
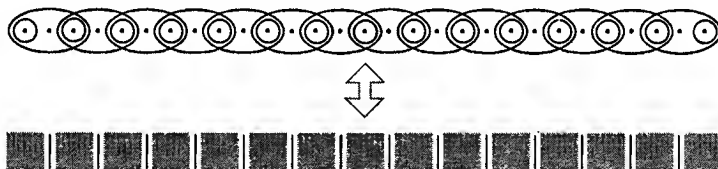
FIGURE 1.20: The digital line models the line of pixels and boundaries between them.

The digital line is not Hausdorff since there is no pair of disjoint open sets U, V with $n \in U$ and $n + 1 \in V$.

Next, we introduce the digital plane. As with the digital line, the idea is to have an array of single-point open sets model the pixels in the digital image display and also to have additional points representing the boundaries between the pixels. We start with $\mathbb{Z} \times \mathbb{Z}$, the product of two copies of the integers. For each point (m, n) we have a basis element $B(m, n)$ defined as follows:

$$B(m, n) = \begin{cases} \{(m, n)\} & \text{if } m \text{ and } n \text{ are odd,} \\ \{(m + a, n) \mid a = -1, 0, 1\} & \text{if } m \text{ is even and } n \text{ is odd,} \\ \{(m, n + b) \mid b = -1, 0, 1\} & \text{if } m \text{ is odd and } n \text{ is even,} \\ \{(m + a, n + b) \mid a, b = -1, 0, 1\} & \text{if } m \text{ and } n \text{ are even.} \end{cases}$$

The basis elements $B(m, n) = \{(m, n)\}$, where m and n are both odd, are the single-point open sets representing the pixels in the digital image display. In Figure 1.21 we show some of the basis elements for the digital plane topology.

The collection $\mathcal{B}_P = \{B(m, n) \mid (m, n) \in \mathbb{Z} \times \mathbb{Z}\}$ is a basis for a topology on $\mathbb{Z} \times \mathbb{Z}$. (See Exercise 1.38.) The resulting topological space is called the **digital plane**. In this topology, the single-point sets $\{(m, n)\}$ are closed sets if
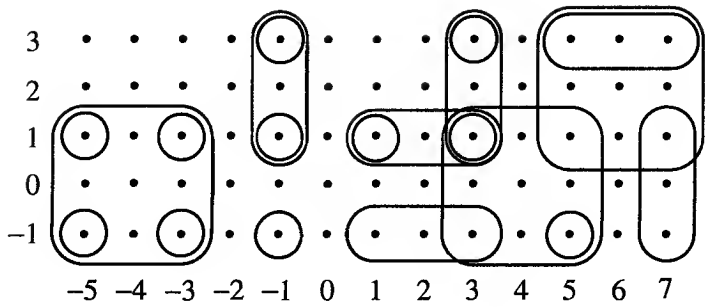
FIGURE 1.21: Some basis elements for the digital plane.

m and n are both even. (See Exercise 1.39.) Furthermore, like the digital line, the digital plane is not a Hausdorff space.

As illustrated in Figure 1.22, we view the digital plane as a set of

(i) Open pixels corresponding to the points (m, n) where m and n are both odd,

(ii) Vertical and horizontal boundaries between the pixels corresponding to the points (m, n) where either m or n is even and the other is odd,

(iii) Closed corner boundary points corresponding to the points (m, n) where both m and n are even.
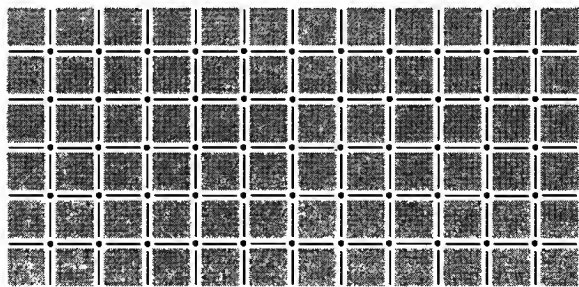
FIGURE 1.22: The digital plane models a rectangular array of pixels and the boundaries between them.

We explore the digital plane further in Section 11.3.

Phenotype Spaces

The genotype–phenotype relationship is of fundamental importance in biology. The genotype is internally coded, inheritable information possessed by all living organisms, while the phenotype is the physical realization of that information. For example, the collection of genes responsible for eye color in a particular individual is a genotype. The observable eye coloration in

the individual is the corresponding phenotype. In this section, we examine a model of evolutionary proximity established by defining a topology on a set of phenotypes. Molecular biologists propose this model as a means for formally defining continuous and discontinuous evolutionary change, providing a mathematical framework for understanding evolutionary processes. (See [Fon1], for example.)

We demonstrate the idea of a phenotype space via an example involving ribonucleic acid (RNA) molecules. Strands of RNA are formed from smaller molecules, called nucleotides, that bond together to make strong, flexible chains. There are four different nucleotides in RNA: guanine (G), cytosine (C), adenine (A) and uracil (U). Nonadjacent nucleotide pairs undergo additional (weaker) bonding, contorting the chain into a more complicated folded arrangement, as illustrated in Figure 1.23.
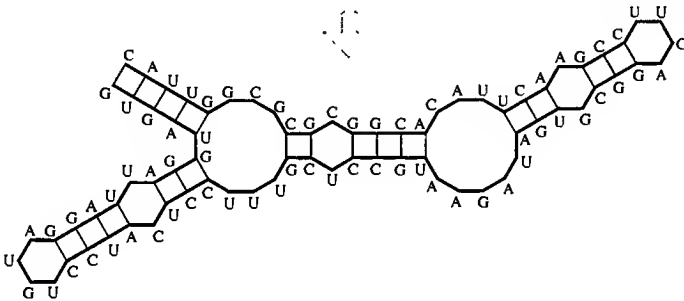
FIGURE 1 23: RNA is a folded and bonded chain of nucleotides.

Pairings of guanine with cytosine and adenine with uracil contribute the most to folding, though guanine and uracil can also pair. In theory, there are many ways that a particular nucleotide chain can fold and bond, but only the most energetically favorable of these are likely to occur. For simplicity, we assume that this bonding occurs in a unique way.

Here, we consider only relatively short nucleotide chains to illustrate the structures and the topological model. RNA molecules in living cells can be anywhere from tens to thousands of nucleotides in length.

For a particular RNA molecule, the associated unfolded nucleotide chain is called the **primary structure**. We represent the nucleotide chain by a sequence of C's, G's, A's and U's, called the **genotype sequence** for the molecule.

Given a genotype sequence, its **bonding diagram** is an unlabeled diagram depicting the bonding that occurs in the resulting RNA molecule. In Figure 1.24 we illustrate the bonding diagram for the RNA molecule in Figure 1.23. The dot on the diagram refers to the location of the first nucleotide in the sequence, and the rest of the sequence wraps counterclockwise around the perimeter of the diagram. The bonding diagram is the phenotype in our model; it is also referred to as the **RNA shape** or **secondary structure** for the RNA molecule. The phenotype spaces we introduce are sets of RNA shapes on which a topology is defined.
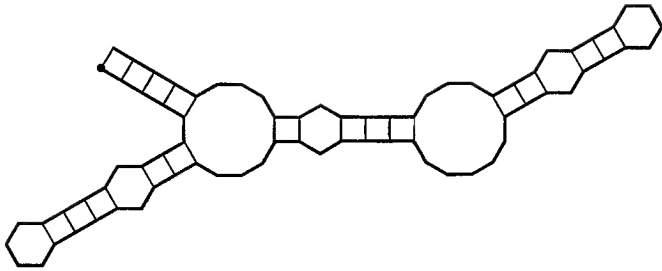
FIGURE 1 24: The bonding diagram for an RNA molecule.

EXAMPLE 1.19. Consider the following three genotype sequences:

1. GGGCAGUCUC CCGGCGUUUA AGGGAUCCUG AACUUCGUCG
    CUCCCAUCCA AUCAGUCCGC CUCACGGAUG GAGUUG

2. GGGCAGUCUC CCGGCGUUUA AGGAAUCCUG AACUUCGUCG
    CUCCCAUCCA AUCAGUCCGC CUCACGGAUG GAGUUG

3. GGGCAGUCUC CCGGCCUUUA AGGGAUCCUG AACUUCGUCG
    CUCCCAUCCA AUCAGUCCGC CUCACGGAUG GAGUUG

The bonding diagrams for these three genotype sequences are shown in Figure 1.25.
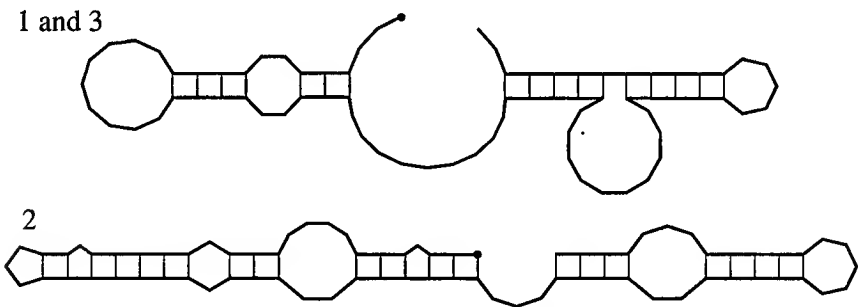
FIGURE 1.25: Bonding diagrams for the nucleotide sequences.

Sequences 1 and 3 are identical except for the 16th entry, and the bonding diagram is the same for both. On the other hand, sequence 2 differs from sequence 1 in only one entry, the 24th, but the corresponding bonding diagrams are very different.

In our model there is a unique RNA shape associated to each genotype sequence. This relationship is not one-to-one since multiple genotype sequences can result in the same RNA shape, as we see in the previous example. An important aspect of this theory is the notion that some single-entry changes in the genotype sequence completely alter the bonding diagram while other changes do not.

DEFINITION 1.20. *The set of all genotype sequences that result in a particular RNA shape s is called the **neutral network** of s and is denoted N(s).*

EXAMPLE 1.20. Consider the genotype sequences of length 10 made up of guanine (G) and cytosine (C) only. There are $2^{10} = 1024$ possible genotype sequences, and, upon folding and bonding, they result in eight different RNA shapes. The RNA shapes are illustrated in Figure 1.26, along with the number of different genotype sequences associated to each (i.e., the size of $N(s)$ for each s). We denote this set of RNA shapes by GC_{10}.
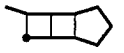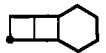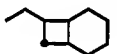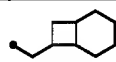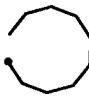
GC_{10}					
S_1	105	S_4	26	S_7	47
S_2	128	S_5	80		
S_3	137	S_6	70	S_8	431

FIGURE 1.26: The set GC_{10}.

In a general biological setting genetic mutation is a random process by which the genotype changes. As a result of the many-to-one nature of the genotype–phenotype mapping, many genetic mutations do not alter the resulting phenotype. On the other hand, there are occasional mutations that result in a new phenotype, and then the process of natural selection determines whether or not the new phenotype persists.

Now, suppose that we have a specific genotype sequence, and let r be its corresponding RNA shape. In this setting, a genetic mutation is a random change in the entries in the genotype sequence. As these random changes occur, we may remain within the neutral network of r, "drifting" through genotype sequences that all result in RNA shape r. At some point, however, a change in one entry in a genotype sequence might take us out of the neutral network of r into that of another RNA shape s. We say that r has mutated to s. (See Figure 1.27.)

Given RNA shapes r and s, we would like to know how likely it is for r to mutate to s as a result of a single-entry change in a genotype sequence in the neutral network of r. We can define and quantify a probability to make this precise. We will use this probability to define phenotype-space topologies on sets of RNA shapes.

Before defining and using the probability, however, we introduce several quantities that play a role in its definition. By a **point mutation** we mean a mutation from one genotype sequence to another obtained by changing a single entry in the sequence. In Example 1.19 sequence 2 is obtained by a point mutation from sequence 1, and vice versa (similarly for sequences 1 and 3).
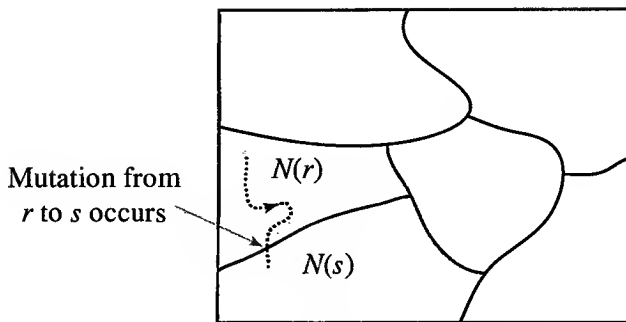
FIGURE 1 27: Genetic mutation from secondary structure r to secondary structure s.

For RNA shapes r and s, let $m_{r,s}$ be the number of point mutations that change a sequence in $N(r)$ to a sequence in $N(s)$ and thus result in a mutation from r to s. Note that $m_{r,s} = m_{s,r}$ because each point mutation changing a sequence in $N(r)$ to one in $N(s)$ has a corresponding inverse point mutation that changes a sequence in $N(s)$ to one in $N(r)$. Let $m_{r,*}$ be the number of point mutations that change a sequence in $N(r)$ to a sequence in any other neutral network. We can think of $m_{r,*}$ as the number of point mutations that take us out of $N(r)$.

DEFINITION 1.21. *The mutation probability,* $p_{r,s}$*, is defined by*

$$p_{r,s} = \frac{m_{r,s}}{m_{r,*}}.$$

Even though $m_{r,s} = m_{s,r}$, it need not be the case that $p_{r,s} = p_{s,r}$ since the values of $m_{r,*}$ and $m_{s,*}$ might be different. For example, if $m_{r,*}$ is greater than $m_{s,*}$, then there are more point mutations out of $N(r)$ than there are out of $N(s)$, and therefore the proportion of mutations $\frac{m_{r,s}}{m_{r,*}}$ is smaller than $\frac{m_{s,r}}{m_{s,*}}$. We will see this asymmetry reflected in the phenotype-space topologies—while RNA shape s might be close to RNA shape r, r need not be close to s.

Because of the asymmetry in these probabilities, a distance function cannot provide a means of determining the proximity of two RNA shapes. A distance function (or metric, as discussed in Chapter 5) is necessarily symmetric; the distance between r and s must equal the distance between s and r. Since a distance function will not serve for this purpose, a topological space is a natural alternative to consider.

EXAMPLE 1.21. In order to clarify how the asymmetry arises in the mutation probabilities, in this example we consider probabilities that are defined like the mutation probabilities, but in a very different setting. Consider the following scenario. In Figure 1.28 we show the six New England states and a table of transition probabilities (defined subsequently) between them. For a pair of states R and S, we define $B_{R,S}$ to be the length of the border between the two states. Here we have $B_{S,R} = B_{R,S}$ for each pair of states R and S.

	CT	MA	ME	NH	RI	VT
CT	—	0.6	0	0	0.4	0
MA	0.3	—	0	0.35	0.2	0.15
ME	0	0	—	1	0	0
NH	0	0.2	0.4	—	0	0.4
RI	0.6	0.4	0	0	—	0
VT	0	0.2	0	0.8	0	—
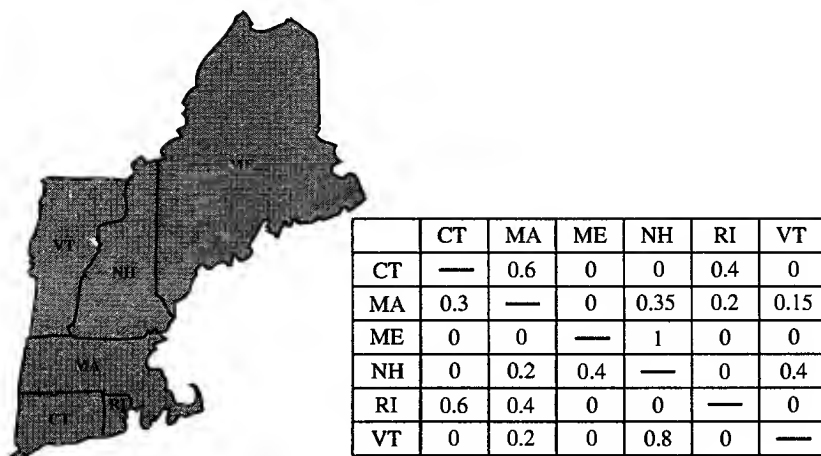
FIGURE 1.28: Transition probabilities between the New England states.

For an individual state R, we define $B_{R,*}$ to be the total length of the border between R and all of the other New England states. We set the transition probability $P_{R,S}$ equal to $B_{R,S}/B_{R,*}$ and interpret this as the probability that a random step out of state R into another New England state lands in state S. The probability $P_{R,S}$ is not symmetric. Since the only state bordering Maine is New Hampshire, $P_{ME,NH} = 1$, but $P_{NH,ME} \neq 1$ since New Hampshire also borders Massachusetts and Vermont. This example is explored further in Exercise 1.44.

For an explicit example, we consider a topology on GC_{10}, the set of RNA shapes of genotypes of length 10. The forthcoming process used in defining a topology from mutation probabilities carries over to RNA shapes associated with genotype sequences of any fixed length.

We do not discuss the details of how the numbers of mutations $m_{r,s}$ and $m_{r,*}$ are determined for the RNA shapes in GC_{10}. For our starting point, we use the associated mutation probabilities in Figure 1.29, which are derived from numbers of mutations as previously described.

	s_1	s_2	s_3	s_4	s_5	s_6	s_7	s_8
s_1	—	0.13	0.15	0.08	0.07	0.09	0.04	0.44
s_2	0.11	—	0.15	0	0.11	0.18	0.05	0.40
s_3	0.12	0.15	—	0.03	0.09	0.06	0.05	0.50
s_4	0.29	0	0.14	—	0.07	0.09	0.06	0.35
s_5	0.08	0.15	0.12	0.02	—	0.08	0.08	0.47
s_6	0.12	0.29	0.09	0.03	0.09	—	0.03	0.35
s_7	0.08	0.12	0.13	0.03	0.13	0.05	—	0.46
s_8	0.18	0.19	0.24	0.04	0.15	0.11	0.09	—

FIGURE 1.29· The mutation probabilities for GC_{10}.

In Figure 1.29 the entry in the ith row and jth column is the probability that a point mutation out of the neutral network of s_i results in a sequence in the neutral network of s_j.

The notion of proximity that we define on the set of RNA shapes is based on the likelihood of a mutation from one RNA shape to another. Since there are eight RNA shapes in GC_{10}, each can potentially mutate to seven others. Hence if $p_{i,j} > 1/7$, we think of s_i as having more than the average likelihood of mutating to s_j.

For each $i = 1, \ldots, 8$, define $R_i = \{s_i\} \cup \{s_j \mid p_{i,j} > 1/7\}$. Thus R_i consists of s_i along with all of the RNA shapes to which s_i has more than the average likelihood of mutating. The collection $\mathcal{R}_{1/7} = \{R_i\}_{i=1}^8$ is not itself a topology, but we extend it to one, defining $\mathcal{T}_{1/7}$ to be the minimal topology on GC_{10} containing $\mathcal{R}_{1/7}$. The topology $\mathcal{T}_{1/7}$ is generated by a basis formed by taking finite intersections of the sets in $\mathcal{R}_{1/7}$. (See Exercise SE 1.20.) The resulting topological space is referred to as a **phenotype space**.

A topology on a finite set has a unique minimal basis that generates the topology. (See Exercises SE 1.23 and SE 1.24.) If for each s_i we take all of the sets R_j that contain s_i, and let B_i be their intersection, then the collection $\mathcal{B}_{1/7} = \{B_i\}_{i=1}^8$ is the minimal basis for $\mathcal{T}_{1/7}$. (See Exercise SE 1.24.)

From the probability table for GC_{10}, it is easy to determine this minimal basis $\mathcal{B}_{1/7}$. To begin, as shown in Figure 1.30, we put a check mark in each diagonal table entry and in each table entry where $p_{i,j} > 1/7$.

	s_1	s_2	s_3	s_4	s_5	s_6	s_7	s_8
s_1	✓		✓					✓
s_2		✓	✓			✓		✓
s_3		✓	✓					✓
s_4	✓			✓				✓
s_5		✓			✓			✓
s_6		✓				✓		✓
s_7							✓	✓
s_8	✓	✓	✓		✓			✓

FIGURE 1.30: The GC_{10} probability-table entries with $p_{i,j} > 1/7$.

For each i, the checkmarks in row i correspond to the elements in R_i. For example, $R_1 = \{s_1, s_3, s_8\}$ and $R_2 = \{s_2, s_3, s_6, s_8\}$. To determine the basis element B_i we take the intersection of the rows that contain a check mark in the ith column. For example, the s_6 column is checked in the second and sixth rows. Intersecting R_2 and R_6 results in $B_6 = \{s_2, s_6, s_8\}$. The basis $\mathcal{B}_{1/7}$ is illustrated in Figure 1.31.

Notice that s_8 is close to each s_j in the sense that the basis element B_j contains s_8 and therefore so does every neighborhood of s_j. This reflects the fact that each s_j mutates to s_8 with a relatively high probability. On the other hand, B_8 contains only s_8, and therefore no other s_j is as close to s_8 as s_8 is

B_1		
B_2		
B_3		
B_4		
B_5		
B_6		
B_7		
B_8		

FIGURE 1.31: The basis $\mathcal{B}_{1/7}$ for the topology $\mathcal{T}_{1/7}$.

to it. While it is a topological consequence that $B_8 = \{s_8\}$, in this particular example having $B_8 = \{s_8\}$ nicely reflects the fact that the neutral network of s_8 is large in comparison to the neutral networks of the other s_j's and therefore mutation out of the neutral network of s_8 is less frequent than mutation into it.

We used the average probability $1/7$ as a threshold in constructing $\mathcal{T}_{1/7}$. That is, we defined sets $R_i = \{s_i\} \cup \{s_j \mid p_{i,j} > 1/7\}$, and we set $\mathcal{T}_{1/7}$ equal to the minimal topology on GC_{10} containing all of the sets R_i. We could work with other probability thresholds and possibly obtain different topologies on the same set of RNA shapes. For example, if we choose $1/10$ as our threshold and let $\mathcal{R}_{1/10}$, $\mathcal{T}_{1/10}$, and $\mathcal{B}_{1/10}$ be defined in the same manner as the corresponding collections above, then the basis $\mathcal{B}_{1/10}$ appears as in Figure 1.32. (See Exercise 1.42.)

Researchers hope that establishing a notion of proximity using topology will help to clarify the biological processes underlying the evolution of living

B_1			
B_2			
B_3			
B_4			
B_5			
B_6			
B_7			
B_8			

FIGURE 1.32: The basis $\mathcal{B}_{1/10}$ for the topology $\mathcal{T}_{1/10}$.

organisms. The simple model presented here, using RNA shapes, can carry over to more general and complex genotype–phenotype systems. Proximity is an important component of continuity, in both real-world and topological terms. (See Chapter 4). Consequently, phenotype spaces provide an appropriate setting for modeling and investigating continuous and discontinuous evolutionary change.

Exercises for Section 1.4

1.38. Show that the collection $\mathcal{B}_P$ of subsets of $\mathbb{Z} \times \mathbb{Z}$, used in defining the digital plane, is a basis for a topology on $\mathbb{Z} \times \mathbb{Z}$.

1.39. For each point (m, n) in the digital plane, determine the smallest closed set containing (m, n). (There are four cases to consider: m and n both odd, m odd and n even, m even and n odd, and m and n both even.)

1.40. Suppose our genotypes are four-element sequences made up of the letters N and S, and we have three phenotypes with neutral networks:
$$N_1 = \{(N, N, N, N), (N, N, N, S), (N, N, S, N), (N, N, S, S)\}$$
$$N_2 = \{(N, S, N, N), (N, S, N, S), (N, S, S, N), (N, S, S, S),$$
$$(S, N, N, N), (S, N, N, S)\}$$
$$N_3 = \{(S, N, S, N), (S, N, S, S), (S, S, N, N), (S, S, N, S),$$
$$(S, S, S, N), (S, S, S, S)\}$$
Determine the mutation probabilities $p_{1,2}$, $p_{1,3}$, $p_{2,3}$, $p_{2,1}$, $p_{3,1}$, and $p_{3,2}$.

1.41. Suppose our genotypes are two-element sequences made up of the letters R, D, and F, and we have three phenotypes with neutral networks:
$$N_1 = \{(R, R)\}$$
$$N_2 = \{(R, D), (F, R), (F, F)\}$$
$$N_3 = \{(D, R), (F, D), (D, D), (R, F), (D, F)\}$$
Determine the mutation probabilities $p_{1,2}$, $p_{1,3}$, $p_{2,3}$, $p_{2,1}$, $p_{3,1}$, and $p_{3,2}$.

1.42. (a) Show the steps in determining $\mathcal{B}_{1/7}$ on GC_{10} from the probability table in Figure 1.29.
(b) Determine $\mathcal{B}_{1/10}$ on GC_{10} from the probability table in Figure 1.29.

1.43. What is the minimum value of p such that $\mathcal{T}_p$ on GC_{10} is the discrete topology? Justify your result.

1.44. Consider the six New England states and the associated probability table in Figure 1.28. Let $NE = \{CT, MA, ME, NH, RI, VT\}$.
(a) Determine the minimal basis for the topology $\mathcal{T}_0$ on NE. In this topology we are considering pairs of states, R and S, where a random step out of state R has a nonzero probability of landing in state S (that is, a step out of R can land in S).
(b) Determine the minimal basis for the topology $\mathcal{T}_{1/3}$ on NE.

Chapter 2

Interior, Closure, and Boundary

In this chapter, we consider some important sets associated with the subsets of a topological space. In particular, we introduce the interior and closure of a set in Section 2.1, the limit points of a set in Section 2.2, and the boundary of a set in Section 2.3. In Section 2.4, we close this chapter with a topological modeling application in geographic information systems.

2.1   Interior and Closure of Sets

An arbitrary subset A of a topological space might be neither open nor closed. However, it is often useful to associate a related open set or a related closed set to A. In particular, we can sandwich each set A between the largest open set contained in A and the smallest closed set containing A. These sets are known as the interior of A and the closure of A, respectively.

DEFINITION 2.1. *Let A be a subset of a topological space X. The **interior of** A, denoted $\mathring{A}$ or Int(A), is the union of all open sets contained in A. The **closure** of A, denoted $\bar{A}$ or $Cl(A)$, is the intersection of all closed sets containing A.*

Clearly, the interior of A is open and a subset of A, and the closure of A is closed and contains A. Thus we have the aforementioned set sandwich, with A caught between an open set and a closed set: $\mathring{A} \subset A \subset \bar{A}$.

The following properties follow readily from the definition of interior and closure.

THEOREM 2.2. *Let X be a topological space and A and B be subsets of X.*

   (i) If U is an open set in X and $U \subset A$, then $U \subset Int(A)$.

   (ii) If C is a closed set in X and $A \subset C$, then $Cl(A) \subset C$.

   (iii) If $A \subset B$ then $Int(A) \subset Int(B)$.

   (iv) If $A \subset B$ then $Cl(A) \subset Cl(B)$.

   (v) A is open if and only if $A = Int(A)$.

   (vi) A is closed if and only if $A = Cl(A)$.

We prove (i), (iii), and (v) here. For (ii), (iv), and (vi) see Exercise 2.2.

Proof of (i). Suppose that U is an open set in X and $U \subset A$. Since Int(A) is the union of all of the open sets that are contained in A, it follows that U is one of the sets making up this union and therefore is a subset of the union. That is, $U \subset$ Int(A). ∎

Proof of (iii). Since $A \subset B$, Int(A) is an open set contained in B. Part (i) implies that every open set contained in B is contained in Int(B). Therefore Int(A) $\subset$ Int(B). ∎

Proof of (v). If $A = $ Int(A), then A is an open set, since by definition Int(A) is an open set.

Now assume that A is open. We show that $A = $ Int(A). First, Int(A) $\subset A$ by definition of Int(A). Furthermore, since A is an open set contained in A, it follows by (i) that $A \subset $ Int(A). Thus $A = $ Int(A) as we wished to show. ∎

Theorem 2.2(i) indicates that every open set contained in A is contained in the interior of A. In this way Int(A) is the largest open set contained in A. Similarly, Theorem 2.2(ii) indicates that every closed set containing A also contains the closure of A, and thus Cl(A) is the smallest closed set containing A. We will find these perspectives on interior and closure very useful when establishing properties of these sets.

EXAMPLE 2.1. Consider $A = [0, 1)$ as a subset of $\mathbb{R}$ with the standard topology. Then $\mathring{A} = (0, 1)$ and $\bar{A} = [0, 1]$.

EXAMPLE 2.2. Consider $A = [0, 1)$ as a subset of $\mathbb{R}$ with the discrete topology. Then $\mathring{A} = \bar{A} = [0, 1)$.

EXAMPLE 2.3. Consider $A = [0, 1)$ as a subset of $\mathbb{R}$ in the finite complement topology. Here $\mathring{A} = \varnothing$ because there are no nonempty open sets contained in $[0, 1)$. Since A is infinite, and the only infinite closed set in this topology is $\mathbb{R}$, it follows that $\bar{A} = \mathbb{R}$.

EXAMPLE 2.4. Consider $A = [0, 1)$ as a subset of $\mathbb{R}$ in the lower limit topology. Here, $\mathring{A} = A$ since A is an open set. Note that $\mathbb{R} - [0, 1) = (-\infty, 0) \cup [1, \infty)$ is an open set in $\mathbb{R}$ with the lower limit topology. Therefore $[0, 1)$ is closed in the lower limit topology, implying that $\bar{A} = A$ as well.

IMPORTANT NOTE: *As Examples 2.1–2.4 demonstrate, the interior and closure of a set A depend on the topology on the set X containing A, not just on the set A.*

EXAMPLE 2.5. Consider the set of rational numbers $\mathbb{Q}$ in $\mathbb{R}$ with the standard topology. We claim that $\mathring{\mathbb{Q}} = \varnothing$. Assume it is not, and suppose that U is a nonempty open set contained in $\mathbb{Q}$. Let x be an element in U. Then there is an open interval (a, b) such that $x \in (a, b) \subset U \subset \mathbb{Q}$. But between every pair of real numbers there is an irrational number. Thus every interval contains elements of $\mathbb{R} - \mathbb{Q}$, and therefore so does U. This is a contradiction; hence $\mathring{\mathbb{Q}} = \varnothing$.

While the interior of $\mathbb{Q}$ contains nothing, if we take the closure we get everything; specifically, $\bar{\mathbb{Q}} = \mathbb{R}$ (see Exercise 2.6).

DEFINITION 2.3. *A subset A of a topological space X is called **dense** if $Cl(A) = X$.*

By Example 2.5, it follows that $\mathbb{Q}$ is dense in $\mathbb{R}$ with the standard topology.

EXAMPLE 2.6. In the finite complement topology on $\mathbb{R}$, every infinite set is dense. Why? In this topology, the closed sets are either finite sets or $\mathbb{R}$ itself. Therefore $\mathbb{R}$ is the only closed set containing an infinite set. Thus, if A is an infinite subset of $\mathbb{R}$, then $Cl(A) = \mathbb{R}$, implying that A is dense in $\mathbb{R}$.

The following two theorems provide simple means for determining when a particular point y is in the interior or in the closure of a given set A:

THEOREM 2.4. *Let X be a topological space, A be a subset of X, and y be an element of X. Then $y \in Int(A)$ if and only if there exists an open set U such that $y \in U \subset A$.*

> **Proof.** First, suppose that there exists an open set U such that $y \in U \subset A$. Then, since U is open and contained in A, it follows that $U \subset Int(A)$. Thus $y \in U$ implies that $y \in Int(A)$.
>
> Next, if $y \in Int(A)$, and we set $U = Int(A)$, it follows that U is an open set such that $y \in U \subset A$. ∎

THEOREM 2.5. *Let X be a topological space, A be a subset of X, and y be an element of X. Then $y \in Cl(A)$ if and only if every open set containing y intersects A.*

> **Proof.** See Exercise 2.10. ∎

EXAMPLE 2.7. In $\mathbb{R}^2$ with the standard topology, let A be the lollipop that appears on the left in Figure 2.1. Then $Int(A)$ appears as in the middle of the figure, and $Cl(A)$ appears as on the right.
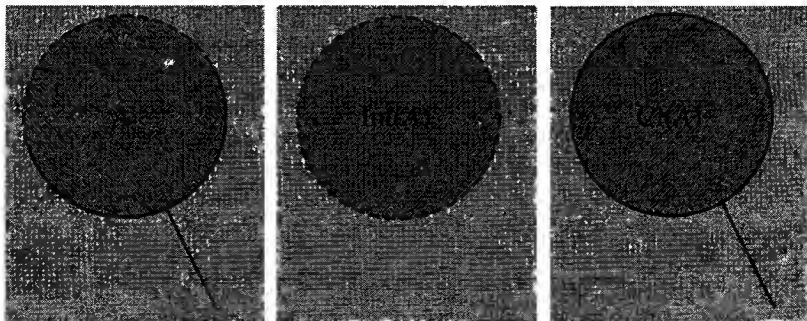
FIGURE 2.1: The interior and closure of a set A in $\mathbb{R}^2$.

The following theorem provides some useful relationships involving interior, closure, intersection, union, and complement:

THEOREM 2.6. *For sets A and B in a topological space X, the following statements hold:*

(i) *$Int(X - A) = X - Cl(A)$.*

(ii) *$Cl(X - A) = X - Int(A)$.*

(iii) *$Int(A) \cup Int(B) \subset Int(A \cup B)$, and in general equality does not hold.*

(iv) *$Int(A) \cap Int(B) = Int(A \cap B)$.*

We prove (i) and (iii) here. For proofs of (ii) and (iv) see Exercise 2.11. In addition, Exercise 2.12 addresses two results involving closure that correspond to (iii) and (iv).

Proof of (i). We prove that $X - Cl(A) \subset Int(X - A)$ and $Int(X - A) \subset X - Cl(A)$. First, note that $Cl(A)$ is closed and contains A, and therefore $X - Cl(A)$ is an open set contained in $X - A$. It follows by Theorem 2.2(i) that $X - Cl(A) \subset Int(X - A)$.

To prove that $Int(X - A) \subset X - Cl(A)$, let $x \in Int(X - A)$ be arbitrary. Note that $Int(X - A)$ is disjoint from A, and therefore x is in an open set that is disjoint from A. By Theorem 2.5, it follows that $x \notin Cl(A)$; hence, $x \in X - Cl(A)$. Thus, $Int(X - A) \subset X - Cl(A)$.

Since we have shown that both $X - Cl(A) \subset Int(X - A)$ and $Int(X - A) \subset X - Cl(A)$ hold, we now have $Int(X - A) = X - Cl(A)$, as we wished to show. ∎

Proof of (iii). Since $Int(A) \subset A$, it follows that $Int(A) \subset A \cup B$. Also, $Int(A)$ is an open set. Similarly, $Int(B)$ is an open set contained in $A \cup B$. Every open set contained in $A \cup B$ is contained in $Int(A \cup B)$. Therefore both $Int(A)$ and $Int(B)$ are contained in $Int(A \cup B)$. Hence their union, $Int(A) \cup Int(B)$, is also contained in $Int(A \cup B)$.

Now we must show that there are cases where $Int(A) \cup Int(B)$ does not equal $Int(A \cup B)$. Take $A = [-1, 0]$ and $B = [0, 1]$ as subsets of $\mathbb{R}$ with the standard topology. Then $Int(A) \cup Int(B) = (-1, 0) \cup (0, 1)$, but $Int(A \cup B) = (-1, 1)$. Thus, in this case $Int(A) \cup Int(B) \neq Int(A \cup B)$. (See Figure 2.2.) ∎
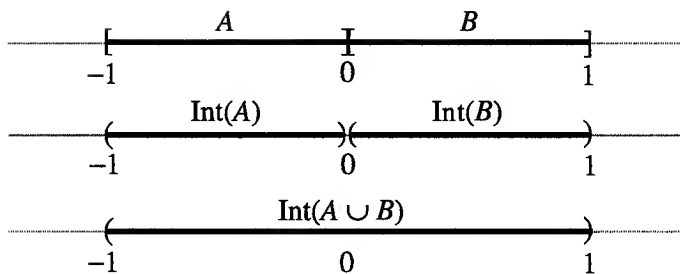
FIGURE 2.2: $\text{Int}(A) \cup Int(B) \neq \text{Int}(A \cup B)$.

EXAMPLE 2.8. Let A be the set of irrational numbers as a subset of $\mathbb{R}$ with the standard topology. We show that A is dense. Note that $A = \mathbb{R} - \mathbb{Q}$ where $\mathbb{Q}$ is the set of rational numbers. Therefore $\text{Cl}(A) = \text{Cl}(\mathbb{R} - \mathbb{Q}) = \mathbb{R} - \text{Int}(\mathbb{Q})$, with the second equality holding by Theorem 2.6(ii). In Example 2.5 we showed that $\text{Int}(\mathbb{Q}) = \varnothing$. Thus $\text{Cl}(A) = \mathbb{R}$, implying that A is dense in $\mathbb{R}$.

Exercises for Section 2.1

2.1. Determine $\text{Int}(A)$ and $\text{Cl}(A)$ in each case.
  (a) $A = (0, 1]$ in the lower limit topology on $\mathbb{R}$.
  (b) $A = \{a\}$ in $X = \{a, b, c\}$ with topology $\{X, \varnothing, \{a\}, \{a, b\}\}$.
  (c) $A = \{a, c\}$ in $X = \{a, b, c\}$ with topology $\{X, \varnothing, \{a\}, \{a, b\}\}$.
  (d) $A = \{b\}$ in $X = \{a, b, c\}$ with topology $\{X, \varnothing, \{a\}, \{a, b\}\}$.
  (e) $A = (-1, 1) \cup \{2\}$ in the standard topology on $\mathbb{R}$.
  (f) $A = (-1, 1) \cup \{2\}$ in the lower limit topology on $\mathbb{R}$.
  (g) $A = \{(x, 0) \in \mathbb{R}^2 \mid x \in \mathbb{R}\}$ in $\mathbb{R}^2$ with the standard topology.
  (h) $A = \{(0, x) \in \mathbb{R}^2 \mid x \in \mathbb{R}\}$ in $\mathbb{R}^2$ with the topology generated by the basis in Exercise 1.19.
  (i) $A = \{(x, 0) \in \mathbb{R}^2 \mid x \in \mathbb{R}\}$ in $\mathbb{R}^2$ with the topology generated by the basis in Exercise 1.19.

2.2. Prove Theorem 2.2, parts (ii), (iv), and (vi): Let X be a topological space and A and B be subsets of X.
  (a) If C is a closed set in X and $A \subset C$, then $\text{Cl}(A) \subset C$.
  (b) If $A \subset B$ then $\text{Cl}(A) \subset \text{Cl}(B)$.
  (c) A is closed if and only if $A = \text{Cl}(A)$.

2.3. For $m < n \in \mathbb{Z}$, let $I_{m,n} = \{m, m + 1, \ldots, n\}$. Determine $\text{Int}(I_{m,n})$ and $\text{Cl}(I_{m,n})$ in the digital line topology. (There are four cases to consider, allowing m to be either even or odd, as well as n.)

2.4. Consider the particular point topology PPX_p on a set X. (See Exercise 1.7.) Determine $\text{Int}(A)$ and $\text{Cl}(A)$ for sets A containing p and for sets A not containing p.

2.5. Consider the excluded point topology EPX_p on a set X. (See Exercise 1.8.) Determine $\text{Int}(A)$ and $\text{Cl}(A)$ for sets A containing p and for sets A not containing p.

2.6. Prove that $Cl(\mathbb{Q}) = \mathbb{R}$ in the standard topology on $\mathbb{R}$.

2.7. Let $B = \{\frac{a}{2^n} \in \mathbb{R} \mid a \in \mathbb{Z}, n \in \mathbb{Z}_+\}$. Show that B is dense in $\mathbb{R}$.

2.8. (a) Show that the set of odd integers is dense in the digital line topology on $\mathbb{Z}$. Is the same true for the set of even integers?

    (b) Which subsets of $\mathbb{Z}$ are dense in the discrete topology on $\mathbb{Z}$?

2.9. In $\mathbb{R}^2$ with the standard topology, prove that $Cl((a, b) \times (c, d)) = [a, b] \times [c, d]$ and $Int([a, b] \times [c, d]) = (a, b) \times (c, d)$.

2.10. **Prove Theorem 2.5:** Let X be a topological space, A be a subset of X, and y be an element of X. Then $y \in Cl(A)$ if and only if every open set containing y intersects A.

2.11. **Prove Tbeorem 2.6, parts (ii) and (iv):** For sets A and B in a topological space X, the following hold:

    (a) $Cl(X - A) = X - Int(A)$.

    (b) $Int(A) \cap Int(B) = Int(A \cap B)$.

2.12. In each case, determine whether the relation in the blank is $\subset$, $\supset$. or $=$. In cases where equality does not hold, provide an example indicating so.

    (a) $Cl(A) \cap Cl(B)$ _____ $Cl(A \cap B)$

    (b) $Cl(A) \cup Cl(B)$ _____ $Cl(A \cup B)$

2.2   Limit Points

In the standard topology on $\mathbb{R}^n$, a limit point of a subset appears as what we might expect; that is, there is a sequence of points in the subset approaching the limit point. This is a very useful concept. However, we make the definition of limit point precise via neighborhoods rather than sequences. In topological spaces that are not quite as well-behaved as $\mathbb{R}^n$, we will see that this allows for some unusual, and perhaps counterintuitive, possibilities.

DEFINITION 2.7. *Let A be a subset of a topological space X. A point x in X is a **limit point** of A if every neighborhood of x intersects A in a point other than x.*

Notice that a limit point x of a set A may or may not lie in the set A. Notice also that in every topology, the point x is not a limit point of the set $\{x\}$.

EXAMPLE 2.9. Consider the set $A = \{\frac{1}{n} \in \mathbb{R} \mid n \in \mathbb{Z}_+\}$ as a subset of $\mathbb{R}$ with the standard topology. It is illustrated in Figure 2.3.
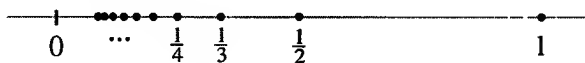
FIGURE 2.3: The set $A = \{\frac{1}{n} \in \mathbb{R} \mid n \in \mathbb{Z}_+\}$.

The point 0 is a limit point of A. Why? If U is an open set containing 0, then there is an interval (a, b) such that $0 \in (a, b) \subset U$. But $(a, b) \cap A \neq \varnothing$

for every such open interval. So $U \cap A \neq \varnothing$. Therefore every neighborhood of 0 intersects A, and since 0 is not in A, the intersection contains points other than 0. It follows that 0 is a limit point of A.

In fact, 0 is the only limit point of A. Given $x \in \mathbb{R} - \{0\}$, we can find an open interval containing x that either is disjoint from A (if $x \notin A$) or intersects A only in x (if $x \in A$). In either case, if $x \neq 0$, then x is not a limit point of A.

EXAMPLE 2.10. Consider $(0, 1]$ as a subset of $\mathbb{R}$ with the standard topology. What are the limit points of $(0, 1]$? Here, every $x \in [0, 1]$ is a limit point of $(0, 1]$. Every neighborhood U of $x \in [0, 1]$ contains an interval (a, b) containing x. Such an interval intersects $(0, 1]$ in a point other than x, and therefore x is a limit point of $(0, 1]$.

Furthermore, if $x \notin [0, 1]$, then there are open intervals containing x that are disjoint from $(0, 1]$. Therefore if $x \notin [0, 1]$, then x is not a limit point of $(0, 1]$.

Thus $[0, 1]$ is the set of limit points of $(0, 1]$.

EXAMPLE 2.11. Consider the rational numbers $\mathbb{Q}$ as a subset of $\mathbb{R}$ with the standard topology. Every $x \in \mathbb{R}$ is a limit point of $\mathbb{Q}$. Why? Given a real number x, an open set U containing x contains an open interval (a, b) that also contains x. But every open interval intersects $\mathbb{Q}$ in infinitely many points, and therefore (a, b) intersects $\mathbb{Q}$ in a point other than x. Hence, x is a limit point of $\mathbb{Q}$.

Limit points provide us with an easy means to find the closure of a set.

THEOREM 2.8. *Let A be a subset of a topological space X, and let A' be the set of limit points of A. Then $Cl(A) = A \cup A'$.*

Proof. We show that $A \cup A' \subset Cl(A)$ and $Cl(A) \subset A \cup A'$.

First we show that $A \cup A' \subset Cl(A)$. Certainly, $A \subset Cl(A)$, so all we need to show is that $A' \subset Cl(A)$. Suppose $x \in A'$. Then every neighborhood of x intersects A. By Theorem 2.5, $x \in Cl(A)$, and it follows that $A' \subset Cl(A)$. Thus $A \cup A' \subset Cl(A)$.

Now consider $Cl(A) \subset A \cup A'$. Suppose $x \in Cl(A)$. Either $x \in A$ or $x \in Cl(A) - A$. In the former case, it follows that $x \in A \cup A'$. Consider the latter case, $x \in Cl(A) - A$. Since $x \in Cl(A)$, Theorem 2.5 implies that every open set containing x intersects A. Since $x \notin A$, such an intersection must contain a point other than x. Thus x is a limit point of A, and it follows that $x \in A \cup A'$. In either case ($x \in A$ or $x \in Cl(A) - A$) we have $x \in A \cup A'$, implying that $Cl(A) \subset A \cup A'$.
∎

COROLLARY 2.9. *A subset A of a topological space is closed if and only if it contains all of its limit points.*

Proof. By Theorem 2.2(vi), A is closed if and only if $A = \text{Cl}(A)$. By Theorem 2.8, $A = \text{Cl}(A)$ if and only if $A = A \cup A'$ where A' is the set of limit points of A. Finally, $A = A \cup A'$ holds if and only if $A' \subset A$. Thus A is closed if and only $A' \subset A$, as we wished to show. ∎

EXAMPLE 2.12. The infinite comb C was introduced in Exercise 1.27 and is shown here in Figure 2.4. The set C consists of the horizontal segment running from $(0, 0)$ to $(1, 0)$ on the x-axis, along with the vertical segments running from $(\frac{1}{2^n}, 0)$ to $(\frac{1}{2^n}, 1)$, for each $n = 0, 1, 2, \dots$.

If we regard C as a subset of $\mathbb{R}^2$ in the standard topology, every point of C is a limit point of C, as are the points in the vertical line segment $Y = \{(0, y) \mid 0 \le y \le 1\}$. As illustrated in Figure 2.4, every neighborhood of each point $(0, y) \in Y$ intersects vertical segments making up C. Besides the points in $C \cup Y$, there are no other limit points of C, so $\text{Cl}(C) = C \cup Y$.
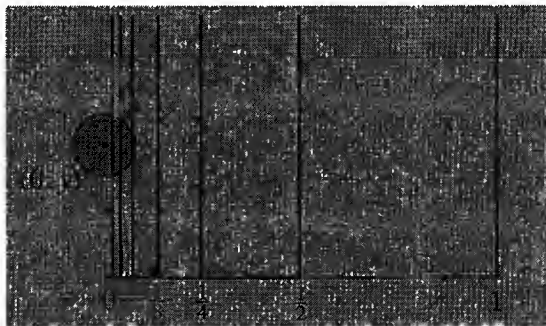
FIGURE 2.4: Points $(0, y)$ with $0 \le y \le 1$ are limit points of the infinite comb.

A concept closely related to limit points is that of a convergent sequence.

DEFINITION 2.10. *In a topological space X, a sequence $(x_1, x_2, \dots)$ **converges to** $x \in X$ if for every neighborhood U of x, there is a positive integer N such that $x_n \in U$ for all $n \ge N$. We say that x is the **limit** of the sequence $(x_1, x_2, \dots)$, and we write*

$$\lim_{n \to \infty} x_n = x.$$

The idea behind a sequence converging to a point x is that, given any neighborhood U of x, the sequence eventually enters and stays in U.

EXAMPLE 2.13. Consider the sequence given by $x_n = \frac{(-1)^n}{n}$ in $\mathbb{R}$, as illustrated in Figure 2.5. In the standard topology on $\mathbb{R}$, the sequence converges to 0 since every neighborhood U of 0 contains an open interval $(-\alpha, \alpha)$ into which the sequence eventually enters and stays. (See Exercise 2.22.)
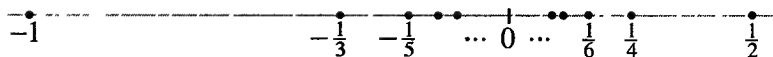
FIGURE 2.5: The sequence $x_n = \frac{(-1)^n}{n}$.

On the other hand, if we consider the lower limit topology on $\mathbb{R}$, this sequence does not converge to 0. Take, for example, the neighborhood $[0, 1)$ of 0. The sequence enters this neighborhood, but never stays in it since every other term jumps out, taking on a value less than 0. In fact, in the lower limit topology on $\mathbb{R}$, this sequence does not converge at all.

In the standard topology on $\mathbb{R}^n$, limit points are limits of sequences, as the following theorem indicates:

THEOREM 2.11. *Let A be a subset of $\mathbb{R}^n$ in the standard topology. If x is a limit point of A, then there is a sequence of points in A that converges to x.*

Proof. See Exercise 2.20. ∎

In a general topological space X, a limit point of a set $B \subset X$ is not necessarily the limit of a sequence in B. (See Exercise 2.23.)

Our intuition might suggest that if a sequence converges to a point, then that point should be unique. That is the case in the standard topology on $\mathbb{R}^n$. But in other topologies this need not be so. For example, in the finite complement topology on $\mathbb{R}$, every sequence with an infinite range converges to every point in $\mathbb{R}$. (See Exercise 2.19.)

Such a situation cannot occur in a Hausdorff space. The next theorem presents another of the convenient properties of a Hausdorff space.

THEOREM 2.12. *If X is a Hausdorff space, then every convergent sequence of points in X converges to a unique point in X.*

Proof. Let X be a Hausdorff space. Suppose that a sequence $(x_1, x_2, \ldots)$ converges to two different points, x and y, in X. Since X is Hausdorff, x and y have disjoint neighborhoods U and V, respectively. The sequence $(x_1, x_2, \ldots)$ converges to x; therefore there exists $N \in \mathbb{Z}_+$ such that $x_n \in U$ for all $n \geq N$. Similarly, there exists $M \in \mathbb{Z}_+$ such that $x_n \in V$ for all $n \geq M$. Thus, if m is greater than both N and M, then $x_m \in U \cap V$, contradicting the fact that U and V are disjoint. Therefore convergent sequences in a Hausdorff space converge to a unique point. ∎

Exercises for Section 2.2

2.13. Determine the set of limit points of A in each case.
   (a) $A = (0, 1]$ in the lower limit topology on $\mathbb{R}$.
   (b) $A = \{a\}$ in $X = \{a, b, c\}$ with topology $\{X, \varnothing, \{a\}, \{a, b\}\}$.
   (c) $A = \{a, c\}$ in $X = \{a, b, c\}$ with topology $\{X, \varnothing, \{a\}, \{a, b\}\}$.
   (d) $A = \{b\}$ in $X = \{a, b, c\}$ with topology $\{X, \varnothing, \{a\}, \{a, b\}\}$.

(e) $A = (-1, 1) \cup \{2\}$ in the standard topology on $\mathbb{R}$.

(f) $A = (-1, 1) \cup \{2\}$ in the lower limit topology on $\mathbb{R}$.

(g) $A = \{(x, 0) \in \mathbb{R}^2 \mid x \in \mathbb{R}\}$ in $\mathbb{R}^2$ with the standard topology.

(h) $A = \{(0, x) \in \mathbb{R}^2 \mid x \in \mathbb{R}\}$ in $\mathbb{R}^2$ with the topology generated by the basis in Exercise 1.19.

(i) $A = \{(x, 0) \in \mathbb{R}^2 \mid x \in \mathbb{R}\}$ in $\mathbb{R}^2$ with the topology generated by the basis in Exercise 1.19.

2.14. For each $n \in \mathbb{Z}_+$, let $B_n = \{n, n+1, n+2, \dots\}$, and consider the collection $\mathcal{B} = \{B_n \mid n \in \mathbb{Z}_+\}$.

(a) Show that $\mathcal{B}$ is a basis for a topology on $\mathbb{Z}_+$.

(b) Show that the topology on X generated by $\mathcal{B}$ is not Hausdorff.

(c) Show that the sequence $(2, 4, 6, 8, \dots)$ converges to every point in $\mathbb{Z}_+$ with the topology generated by $\mathcal{B}$.

(d) Prove that every sequence with infinite range converges to every point in $\mathbb{Z}_+$ with the topology generated by $\mathcal{B}$.

2.15. Determine the set of limit points of $[0, 1]$ in the finite complement topology on $\mathbb{R}$.

2.16. Determine the set of limit points of the single-point set $\{n\}$ in the digital line topology. (The result depends on whether n is odd or even.)

2.17. (a) Let $\mathcal{B} = \{[a, b) \subset \mathbb{R} \mid a, b \in \mathbb{Q} \text{ and } a < b\}$. Show that $\mathcal{B}$ is a basis for a topology on $\mathbb{R}$. The resulting topology is called the **rational lower limit topology** and is denoted $\mathbb{R}_{rl}$.

(b) Determine the closures of $A = (0, \sqrt{2})$ and $B = (\sqrt{2}, 3)$ in $\mathbb{R}_l$ and in $\mathbb{R}_{rl}$.

2.18. Determine the set of limit points of $A = \{\frac{1}{m} + \frac{1}{n} \in \mathbb{R} \mid m, n \in \mathbb{Z}_+\}$ in the standard topology on $\mathbb{R}$.

2.19. Show that if (x_n) is a sequence in $\mathbb{R}$ having infinite range, then (x_n) converges to every point in $\mathbb{R}$ with the finite complement topology on $\mathbb{R}$.

2.20. **Prove Theorem 2.11:** Let A be a subset of $\mathbb{R}^n$ in the standard topology. If x is a limit point of A, then there is a sequence of points in A that converges to x.

2.21. Determine the set of limit points of the set

$$S = \{(x, \sin(\tfrac{1}{x})) \in \mathbb{R}^2 \mid 0 < x \le 1\}$$

as a subset of $\mathbb{R}^2$ in the standard topology. (The closure of S in the plane is known as the **topologist's sine curve**.)

2.22. Consider the sequence defined by $x_n = \frac{(-1)^n}{n}$ in $\mathbb{R}$ with the standard topology.

(a) Prove that every neighborhood of the point 0 contains an open interval $(-\alpha, \alpha)$.

(b) Prove that for each open interval $(-\alpha, \alpha)$, there exists $N \in \mathbb{Z}_+$ such that $x_n \in (-\alpha, \alpha)$ for all $n \ge N$.

2.23. Let $\mathcal{T}$ be the collection of subsets of $\mathbb{R}$ consisting of the empty set and every set whose complement is countable.

(a) Show that $\mathcal{T}$ is a topology on $\mathbb{R}$. (It is called the **countable complement topology**.)

(b) Show that the point 0 is a limit point of the set $A = \mathbb{R} - \{0\}$ in the countable complement topology.

(c) Show that in $A = \mathbb{R} - \{0\}$ there is no sequence converging to 0 in the countable complement topology.

2.3   *The Boundary of a Set*

We have an intuitive idea of what we mean by the boundary of a set: the points that lie close to both the inside and the outside of the set. (See Figure 2.6.) But as we have seen, there are examples and topologies out there that challenge or defy our intuition. For example, what is the boundary of $\mathbb{Q}$ as a subset of $\mathbb{R}$ in the standard topology? Also, what is the boundary of the closed interval $[0, 1]$ as a subset of $\mathbb{R}$ in the finite complement topology? The definition of the boundary of a set has to be chosen carefully and understood clearly for its proper use as a central concept in topology.
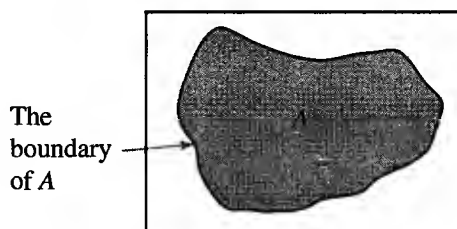
The boundary of A

FIGURE 2.6: The boundary of a set.

DEFINITION 2.13. *Let A be a subset of a topological space X. The **boundary** of A, denoted ∂A, is the set $\partial A = Cl(A) - Int(A)$.*

EXAMPLE 2.14. Let $A = [-1, 1]$ in the standard topology on $\mathbb{R}$. As illustrated in Figure 2.7, we have $Cl(A) = [-1, 1]$ and $Int(A) = (-1, 1)$, and therefore $\partial A = \{-1, 1\}$.

$$A = Cl(A)$$
$$-1 \qquad\qquad 1$$
$$Int(A)$$
$$-1 \qquad\qquad 1$$
$$\partial(A) \qquad\qquad \partial(A)$$
$$-1 \qquad\qquad 1$$
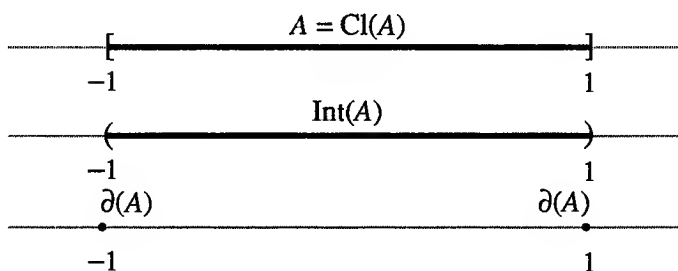
FIGURE 2.7: Determining the boundary of $[-1, 1]$ in $\mathbb{R}$.

EXAMPLE 2.15. Let A be the subset of the plane appearing on the far left in Figure 2.8. Then, as pictured, $Cl(A)$ is a closed ball, $Int(A)$ is an open ball, and ∂A is a circle.

The next theorem gives conditions that help us to determine whether or not a point x lies in the boundary of a set A.
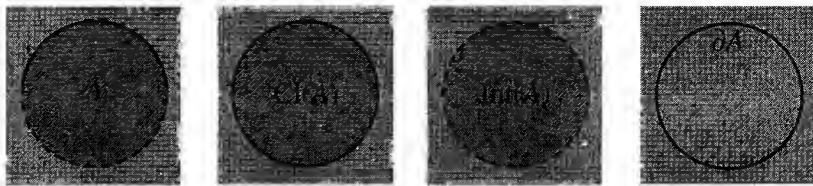
FIGURE 2.8: Determining the boundary of A in the plane.

THEOREM 2.14. *Let A be a subset of a topological space X and let x be a point in X. Then $x \in \partial A$ if and only if every neighborhood of x intersects both A and $X - A$.*

> ***Proof.*** To begin, suppose that $x \in \partial A$. Then $x \in \mathrm{Cl}(A)$ and $x \notin \mathrm{Int}(A)$. Since $x \in \mathrm{Cl}(A)$, it follows that every neighborhood of x intersects A. Furthermore, since $x \notin \mathrm{Int}(A)$, it follows that every neighborhood of x is not a subset of A and therefore intersects $X - A$. Thus, every neighborhood of x intersects A and $X - A$.
>
> Now suppose that every neighborhood of x intersects A and $X - A$. It follows that $x \in \mathrm{Cl}(A)$ and $x \in \mathrm{Cl}(X - A)$. By Theorem 2.6 implies that $\mathrm{Cl}(X - A) = X - \mathrm{Int}(A)$, and therefore $x \notin \mathrm{Int}(A)$. Thus, $x \in \mathrm{Cl}(A)$ and $x \notin \mathrm{Int}(A)$; that is, $x \in \mathrm{Cl}(A) - \mathrm{Int}(A) = \partial A$. ∎

Here are some quick facts about the boundary of a set A. They all follow readily from the definition.

THEOREM 2.15. *Let A be a subset of a topological space X. Then the following statements about the boundary of A hold:*

(i) ∂A *is closed.*

(ii) $\partial A = Cl(A) \cap Cl(X - A)$.

(iii) $\partial A \cap Int(A) = \varnothing$.

(iv) $\partial A \cup Int(A) = Cl(A)$.

(v) $\partial A \subset A$ *if and only if A is closed.*

(vi) $\partial A \cap A = \varnothing$ *if and only if A is open.*

(vii) $\partial A = \varnothing$ *if and only if A is both open and closed.*

Proof. See Exercise 2.28. ∎

EXAMPLE 2.16. Consider $\mathbb{Q}$ in the standard topology on $\mathbb{R}$. Since $\mathrm{Cl}(\mathbb{Q}) = \mathbb{R}$, and $\mathrm{Int}(\mathbb{Q}) = \varnothing$, it follows that $\partial \mathbb{Q} = \mathbb{R}$.

Therefore the whole real line $\mathbb{R}$ is the boundary of the rational numbers. That makes sense—every real number is arbitrarily close to the set of rational numbers and to its complement, the set of irrational numbers.

EXAMPLE 2.17. Consider the vertical interval $A = \{0\} \times [-1, 1]$ as a subset of $\mathbb{R}^2$ with the standard topology. Viewing the interval this way, as a subset of $\mathbb{R}^2$ rather than as a subset of $\mathbb{R}$, makes a huge difference in how its boundary appears. Here, $\text{Int}(A) = \varnothing$ and $\text{Cl}(A) = A$; therefore $\partial A = A$.

EXAMPLE 2.18. Let $A = [-1, 1]$ in $\mathbb{R}$ with the discrete topology. Here $\text{Int}(A) = [-1, 1]$ and $\text{Cl}(A) = [-1, 1]$, so $\partial A = \varnothing$.

Alternatively, we could observe that $[-1, 1]$ is both open and closed in the discrete topology, and therefore its boundary is empty by Theorem 2.15(vii).

EXAMPLE 2.19. Let $A = [-1, 1]$ as a subset of $\mathbb{R}_l$, the real line with the lower limit topology. In this case, $\text{Cl}(A) = [-1, 1]$ and $\text{Int}(A) = [-1, 0)$, so $\partial A = \{-1\}$.

The previous examples show that, as with interior and closure, the boundary of a set A depends on the topology on the set X containing A, not just on A itself.

Exercises for Section 2.3

2.24. Determine ∂A in each case.
   (a) $A = (0, 1]$ in the lower limit topology on $\mathbb{R}$.
   (b) $A = \{a\}$ in $X = \{a, b, c\}$ with topology $\{X, \varnothing, \{a\}, \{a, b\}\}$.
   (c) $A = \{a, c\}$ in $X = \{a, b, c\}$ with topology $\{X, \varnothing, \{a\}, \{a, b\}\}$.
   (d) $A = \{b\}$ in $X = \{a, b, c\}$ with topology $\{X, \varnothing, \{a\}, \{a, b\}\}$.
   (e) $A = (-1, 1) \cup \{2\}$ in the standard topology on $\mathbb{R}$.
   (f) $A = (-1, 1) \cup \{2\}$ in the lower limit topology on $\mathbb{R}$.
   (g) $A = \{(x, 0) \in \mathbb{R}^2 \mid x \in \mathbb{R}\}$ in $\mathbb{R}^2$ with the standard topology.
   (h) $A = \{(0, x) \in \mathbb{R}^2 \mid x \in \mathbb{R}\}$ in $\mathbb{R}^2$ with the topology generated by the basis in Exercise 1.19.
   (i) $A = \{(x, 0) \in \mathbb{R}^2 \mid x \in \mathbb{R}\}$ in $\mathbb{R}^2$ with the topology generated by the basis in Exercise 1.19.

2.25. (a) For $m < n \in \mathbb{Z}$, let $I_{m,n} = \{m, m + 1, \ldots, n\}$. Determine $\partial(I_{m,n})$ in the digital line topology. (There are four cases to consider, allowing m to be either even or odd, as well as n.)
   (b) For $n \in \mathbb{Z}$, determine $\partial(\{n\})$ in the digital line topology, considering separately the cases where n is even and n is odd. Discuss how your results for $\partial(\{n\})$ reflect the digital image display structure modeled by the digital line as described in Section 1.4.

2.26. Determine the boundary of each of the following subsets of $\mathbb{R}^2$ in the standard topology:
   (a) $A = \{(x, 0) \in \mathbb{R}^2 \mid x \in \mathbb{R}\}$.
   (b) $B = \{(x, y) \in \mathbb{R}^2 \mid x > 0, y \neq 0\}$.
   (c) $C = \{(\frac{1}{n}, 0) \in \mathbb{R}^2 \mid n \in \mathbb{Z}_+\}$.
   (d) $D = \{(x, y) \in \mathbb{R}^2 \mid 0 \leq x^2 - y^2 < 1\}$.

2.27. Determine $\partial([0, 1])$ in $\mathbb{R}$ with the finite complement topology. Justify your result.

2.28. Prove Theorem 2.15: Let A be a subset of a topological space X.

    (a) ∂A is closed.
    (b) $\partial A = \mathrm{Cl}(A) \cap \mathrm{Cl}(X - A)$.
    (c) $\partial A \cap \mathrm{Int}(A) = \varnothing$.
    (d) $\partial A \cup \mathrm{Int}(A) = \mathrm{Cl}(A)$.
    (e) $\partial A \subset A$ if and only if A is closed.
    (f) $\partial A \cap A = \varnothing$ if and only if A is open.
    (g) $\partial A = \varnothing$ if and only if A is both open and closed.

2.4   An Application to Geographic Information Systems

A geographic information system (GIS) is a computer system capable of assembling, storing, manipulating, and displaying geographically referenced data. The data are often used for solving complex planning and management problems. To analyze spatial information, users select data from a GIS by submitting queries. Typical GIS queries incorporate spatial relations to describe constraints about spatial objects to be analyzed or displayed.

For example, in studying wetland protection within a state's recreation areas, a GIS user might ask for a display of all wetlands that lie partially or entirely within the state's parklands. (See Figure 2.9.) The GIS would search for wetland areas and state parkland areas and examine the relationship of each to the other in order to return all wetlands that satisfied the specified requirements.
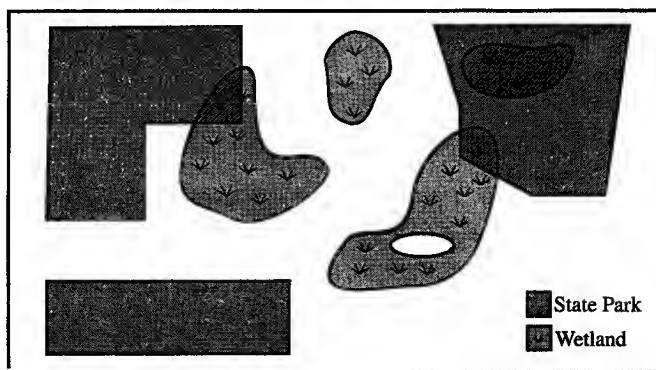
FIGURE 2.9: Which wetlands lie partly or entirely within a state park?

Evidently, in a GIS there is a need to be able to distinguish different ways that land regions can lay in relation to each other. For a mathematician, it often suffices to know if two sets intersect or not, but in a GIS a classification finer than intersect/not-intersect is needed. For example, in Figure 2.10, while sets A and B and sets A' and B' intersect, there is an obvious difference in the nature of their intersections. There is a need to make these distinctions precise.
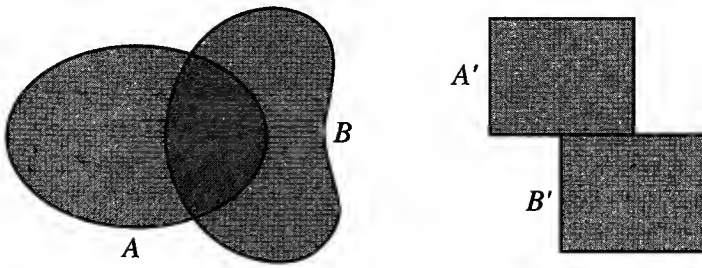
FIGURE 2.10: *A* and *B* intersect in a different manner than do *A'* and *B'*.

Furthermore, in Figure 2.11, while the initial view of sets C and D suggests that they intersect, after zooming in we see that they do not. Information stored in a GIS regarding the relationship between such sets should not depend on how the sets appear in a particular picture. Thus, by storing the information that C and D do not intersect, the GIS ensures that the nature of the relationship between the two sets is independent of potentially misleading pictorial representations of them, particularly those characterized by poor resolution.
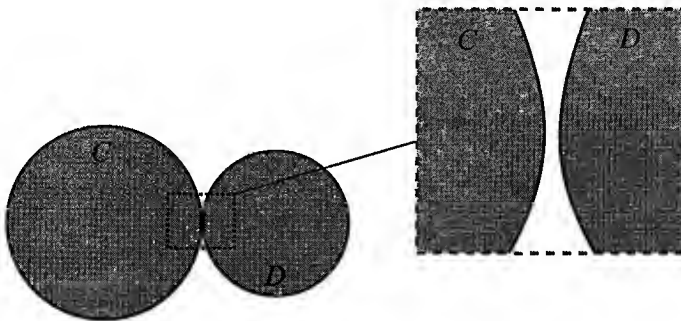
FIGURE 2.11: Sets C and D appear to intersect, but zooming shows they do not.

Let us take another view of the GIS modeling challenge. Computer systems are very good at modeling and manipulating one-dimensional information. One-dimensional information is essentially numerical in nature, and computer systems possess a straightforward, workable model of the real number system, its arithmetic operations, and its order relationship. Computers can easily handle queries about one-dimensional information, such as "If I write out a check for $1000 will I overdraw my account?" and "What is the total population of the six New England states?"

But when the nature of the information moves up a dimension, the modeling challenge becomes considerably more difficult. A model must allow answers to queries such as "In which states does Yellowstone National Park lie?" and "Is Illinois north of Florida?" A system capable of addressing such queries must keep track of the geographic objects, be able to perform operations on them, and be able to precisely define relationships such as "lies in" and "is north of." Furthermore, such a system must match our intuition and understanding of such objects, operations, and relationships. Clearly, this is a significant modeling challenge.

Geographic information systems theory is one area in the field of Spatial Information Science and Engineering, a field that has undergone tremendous growth in recent years. It is a field where topology has provided valuable modeling tools. We present a simple model, originally published in [Ege], that employs topological concepts to define and distinguish relationships between pairs of geographic areas. This model is by no means complete, but it is a start. It has been adopted as a GIS-industry standard for describing the relationships it addresses.

In the model we present here, our interest is in pairs of closed sets A and B in a topological space X, and our goal is to use topological concepts to examine the different ways in which A and B can lie in relation to each other in X.

Given closed sets A and B in X, we consider the four intersections $\partial A \cap \partial B$, $\mathring{A} \cap \mathring{B}$, $\partial A \cap \mathring{B}$, and $\mathring{A} \cap \partial B$, examining whether or not they are empty.

DEFINITION 2.16. *For a set Y, define $C_Y = 0$ if Y is empty and $C_Y = 1$ if Y is not empty. Given closed sets A and B in X, define $\mathbf{I_{A,B}}$, **the intersection value for A and B in X,** by*

$$I_{A,B} = (C_{\partial A \cap \partial B}, C_{\mathring{A} \cap \mathring{B}}, C_{\partial A \cap \mathring{B}}, C_{\mathring{A} \cap \partial B}).$$

Examples of intersection values for pairs of sets in the plane are shown in Figure 2.12.

$(1, 1, 1 ,1)$ $(1, 1, 1, 1)$ $(1, 0, 0, 0)$
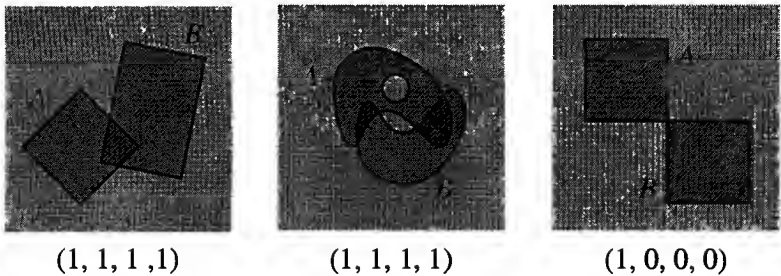
FIGURE 2.12: Intersection values for various pairs of sets in the plane.

Since each of the four entries in the intersection value can be either 0 or 1, there are 16 different possibilities for the intersection value. All 16 can be realized by pairs of closed sets in the plane. In Figure 2.13 we show eight of them. We ask you to find examples of the other eight in Exercise 2.34.

In an application such as a GIS, we may want to restrict the types of sets under consideration. For example, we may wish to model only geographic areas such as the ones shown in Figure 2.9. Thus we would exclude sets such as those shown in Figure 2.13 that have a segment sticking out. Although we might want to allow infinitely thin arcs when we examine borders between countries, we do not want to allow them to represent land areas. How can we eliminate this whole class of possibilities? The next definition helps us do so.

$$(1, 0, 0, 0) \qquad (0, 0, 0, 1) \qquad (1, 1, 0, 1) \qquad (1, 1, 1, 1)$$

$$(0, 0, 0, 0) \qquad (0, 1, 0, 0) \qquad (0, 1, 0, 1) \qquad (1, 0, 1, 1)$$
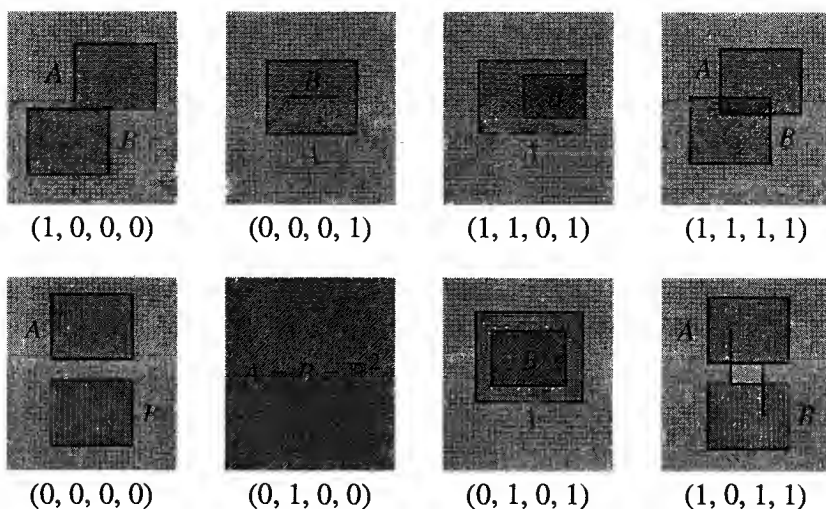
FIGURE 2.13: Pairs of closed sets in the plane realizing eight of the possible intersection values.

DEFINITION 2.17. *A set A for which* $A = Cl(Int(A))$ *is called* **regularly closed**.

For example, closed intervals $[a, b]$ with $a < b$ are regularly closed in $\mathbb{R}$ with the standard topology, since the interior is (a, b) and the closure of the interior is the original set $[a, b]$. The sets appearing on the left in Figure 2.14 are regularly closed subsets of the plane in the standard topology, while those appearing on the right are not.
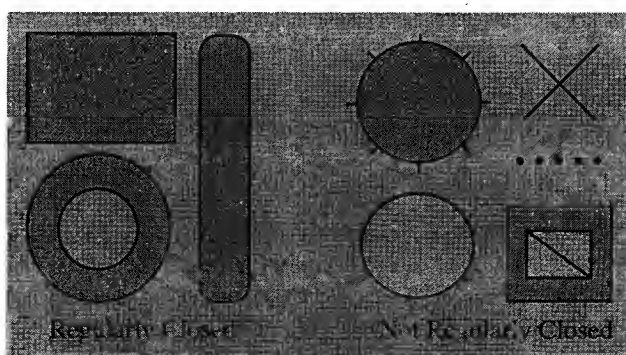
FIGURE 2.14: Closed sets in the plane, some regularly closed, some not.

Regularly closed subsets of the plane have no whiskers attached; all boundary points have interior points arbitrarily close by. (See Exercise 2.35.) Such a feature is a natural requirement in modeling geographic areas. Once we restrict ourselves to regularly closed sets, only some of the 16 intersection values will be possible. We make this explicit via the following theorem and corollary:

THEOREM 2.18. *Let A and B be regularly closed in a topological space X. If $\partial A \cap \mathring{B} \neq \varnothing$, then $\mathring{A} \cap \mathring{B} \neq \varnothing$.*

> **Proof.** First, we prove that $\partial A \subset Cl(\mathring{A})$. Since A is closed, it follows that $\partial A \subset A$. Furthermore, $A = Cl(\mathring{A})$ by the definition of regularly closed. Thus, $\partial A \subset Cl(\mathring{A})$.
>
> Now, to prove that if $\partial A \cap \mathring{B} \neq \varnothing$, then $\mathring{A} \cap \mathring{B} \neq \varnothing$, assume $\partial A \cap \mathring{B} \neq \varnothing$, and let x be an element of this intersection. Since $x \in \partial A \subset Cl(\mathring{A})$, it follows that every open set containing x intersects $\mathring{A}$. But, $\mathring{B}$ is an open set containing x; therefore $\mathring{B}$ intersects $\mathring{A}$. Hence, $\mathring{A} \cap \mathring{B} \neq \varnothing$, as we wished to show.   ∎

The following corollary is now evident:

COROLLARY 2.19. *Intersection values $(1, 0, 1, 0)$, $(0, 0, 1, 0)$, $(1, 0, 0, 1)$, $(0, 0, 0, 1)$, $(1, 0, 1, 1)$, and $(0, 0, 1, 1)$ are not possible for a pair of regularly closed sets A and B in a topological space X.*

Each of the 10 intersection values not excluded by Corollary 2.19 can be realized by pairs of regularly closed sets in the plane. Some are shown in Figure 2.13; for the rest, see Exercise 2.34.

The intersection value provides us with a simple means for distinguishing the different ways two sets can intersect, but it does not distinguish many different configurations. There are, however, situations where the intersection value allows us to draw specific conclusions about the nature of the relationship between two sets. For example, it is easy to see that if $I_{A,B} = (0, 0, 0, 0)$ then A and B are disjoint. With some straightforward assumptions about A and B we obtain a collection of sets for which the intersection value allows other specific conclusions to be drawn about the nature of the configuration of the two sets.

DEFINITION 2.20. *A **planar spatial region** is a nonempty proper subset C of $\mathbb{R}^2$ satisfying the following conditions:*

> *(i) It is regularly closed.*
>
> *(ii) The interior of C cannot be expressed as the union of two disjoint nonempty open sets.*

We have already addressed the significance of being regularly closed. The second condition in the definition of a planar spatial region is related to connectedness, a topological property that is the subject of Chapter 6. This condition requires that the set not have separate interior parts that are, for instance, only accessible from each other by going through the boundary. (See Figure 2.15.)

Closed balls and polygons are examples of planar spatial regions. Results about connectedness are needed to prove that such sets satisfy the second condition for being a planar spatial region. We develop those results in Chapter 6.

In Figure 2.15 we show examples of subsets of the plane that are not planar spatial regions. The set A is not a planar spatial region because it is not regularly
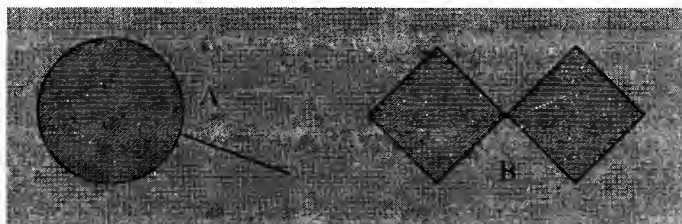
FIGURE 2.15: Sets that are not planar spatial regions.

closed. The set B is not a planar spatial region because its interior is the union of two disjoint nonempty open sets.

The following theorem indicates that, if we restrict ourselves to planar spatial regions, the intersection value enables us to make specific conclusions about the relationship between the sets involved:

THEOREM 2.21. *Let A and B be planar spatial regions. Only the intersection values listed in the left column of the following table are possible for A and B. Furthermore, if the intersection value is as depicted in the left column, then the relationship between A and B is as depicted in the corresponding entry in the right column.*

Intersection Value	Relationship
$(1, 1, 1, 0)$	$A \subset B$
$(0, 1, 1, 0)$	$A \subset \text{Int}(B)$
$(1, 1, 0, 1)$	$B \subset A$
$(0, 1, 0, 1)$	$B \subset \text{Int}(A)$
$(1, 1, 0, 0)$	$A = B$
$(0, 0, 0, 0)$	$A \cap B = \varnothing$
$(1, 0, 0, 0)$	$A \cap B = \partial A \cap \partial B$
$(0, 1, 1, 1)$	$\text{Int}(A) \cap \text{Int}(B) \neq \varnothing,$
$(1, 1, 1, 1)$	$A \not\subset B$, and $B \not\subset A$

The results in the last four rows of the table actually hold for every pair of closed sets A and B in a topological space X. (See Exercise 2.30.) The proofs of the remaining conclusions in this theorem require aspects of connectedness that are established in Chapter 6. We revisit these results in a set of supplementary exercises in Section 6.1.

The result in the fifth row is particularly interesting. Given two planar spatial regions, if we know that their intersection value is $(1,1,0,0)$, then we can conclude that the sets must be equal.

Both of the properties required for a set to be a planar spatial region are needed in Theorem 2.21. If we do not require both properties, then it is possible to find unequal sets A and B for which $(1,1,0,0)$ is the intersection value. (See Exercise 2.31.)

The Open Geospatial Consortium is a group of software developers who have jointly established industry standards for GIS software. They have adopted a topologically based function, expanding on the intersection value, for coding how two geographic areas lie in relation to each other. (See [Ryd].) We will not be specific about what we mean by "geographic areas," but they are examples of planar spatial regions.

Since geographic areas are planar spatial regions, Theorem 2.21 applies. Consequently the Open Geospatial Consortium has adopted a language for describing the relationships between pairs of geographic areas, associating descriptive expressions to the intersection values as shown in the following table. These descriptive expressions cover all possibilities for the relationship between two geographic areas (as a result of Theorem 2.21) and they are mutually exclusive since they are associated to specific intersection values.

Descriptive Expression	Intersection Value	Relationship
A is disjoint from B	$(0, 0, 0, 0)$	$A \cap B \neq \varnothing$
A touches B	$(1, 0, 0, 0)$	$A \cap B = \partial A \cap \partial B$
A equals B	$(1, 1, 0, 0)$	$A = B$
A is within B	$(1, 1, 1, 0)$ or $(0, 1, 1, 0)$	$A \subset B$
A contains B	$(1, 1, 0, 1)$ or $(0, 1, 0, 1)$	$A \supset B$
A overlaps B	$(1, 1, 1, 1)$ or $(0, 1, 1, 1)$	$\text{Int}(A) \cap \text{Int}(B) \neq \varnothing$, $A \not\subset B$, and $B \not\subset A$

Recall the scenario described previously where a GIS user submits a request for a display of all wetlands that lie partially or entirely within a state's parklands. In response to the query, the GIS checks each pair of stored geographic areas consisting of a wetland area A and a state parkland area B and returns all of those pairs for which A equals B, A is within B, A contains B, or A overlaps B.

Exercises for Section 2.4

2.29. Let $A = [a_1, a_2]$ and $B = [b_1, b_2]$ be closed and bounded intervals in $\mathbb{R}$. In each of the following groups determine which of the four possible intersection values can be realized and which cannot for A and B in $\mathbb{R}$. Depict those that can be realized and prove that the remainder cannot.
(a) $(0, 0, 0, 0)$, $(1, 0, 0, 0)$, $(0, 1, 0, 0)$, $(1, 1, 0, 0)$
(b) $(0, 0, 1, 0)$, $(1, 0, 1, 0)$, $(0, 1, 1, 0)$, $(1, 1, 1, 0)$
(c) $(0, 0, 0, 1)$, $(1, 0, 0, 1)$, $(0, 1, 0, 1)$, $(1, 1, 0, 1)$
(d) $(0, 0, 1, 1)$, $(1, 0, 1, 1)$, $(0, 1, 1, 1)$, $(1, 1, 1, 1)$

2.30. Let A and B be closed sets in a topological space X.
(a) Prove that if $I_{A,B} = (0, 0, 0, 0)$, then $A \cap B = \varnothing$.
(b) Prove that if $I_{A,B} = (1, 0, 0, 0)$, then $A \cap B = \partial A \cap \partial B$.
(c) Prove that if $I_{A,B} = (1, 1, 1, 1)$ or $(0, 1, 1, 1)$, then $\text{Int}(A) \cap \text{Int}(B) \neq \varnothing$, $A \not\subset B$, and $B \not\subset A$.

2.31. This exercise demonstrates that if we drop either of the defining conditions for planar spatial regions A and B, then $I_{A,B} = (1, 1, 0, 0)$ need not imply $A = B$.

    (a) Find an example of regularly closed sets A and B in the plane such that $I_{A,B} = (1, 1, 0, 0)$ and $A \neq B$.

    (b) Find an example of closed sets A and B in the plane, each having an interior that is an open ball, such that $I_{A,B} = (1, 1, 0, 0)$ and $A \neq B$. (In Chapter 6 we prove that an open ball in $\mathbb{R}^2$ is connected, meaning it cannot be expressed as the union of two disjoint nonempty open subsets. Therefore a set with an interior that is an open ball satisfies the second condition to be a planar spatial region.)

2.32. Prove that if U is open and $B = \text{Cl}(U)$, then B is regularly closed.

2.33. Prove that a closed rectangle $[a, b] \times [c, d]$ is regularly closed as a subset of the plane.

2.34. Provide examples of pairs of closed sets in the plane that realize the eight intersection values not depicted in Figure 2.13. Use pairs of regularly closed sets wherever possible.

2.35. (a) Provide an example demonstrating that ∂A need not equal $\partial(\text{Int}(A))$.

    (b) Show that $\partial A = \partial(\text{Int}(A))$ for regularly closed sets A.

Creating New Topological Spaces

In this chapter, we look at some methods for creating new topological spaces, given particular topological spaces with which to start. In Section 3.1, we show how a subset of a topological space inherits a topology from the space itself. In Section 3.2, we show how combining topological spaces by taking their product results in a new topological space, the product space. In Sections 3.3 and 3.4, we show how to glue portions of a topological space together to obtain a new space called a quotient space. Finally, in Section 3.5, we introduce configuration spaces and phase spaces, and we examine examples of how they arise in the physical sciences.

Over the course of the chapter we introduce a number of new topological spaces. We continue to use the expression "topologically equivalent" when describing spaces that are indistinguishable as far as topology is concerned. In Chapter 4 we formally define topological equivalence via functions known as homeomorphisms.

3.1   *The Subspace Topology*

Given a subset Y of a topological space X, there is a natural way to define a topology on Y, based on the topology on X.

DEFINITION 3.1. *Let X be a topological space and let Y be a subset of X. Define $\mathcal{T}_Y = \{U \cap Y \mid U \text{ is open in } X\}$. This is called the **subspace topology** on Y and, with this topology, Y is called a **subspace** of X. We say that $V \subset Y$ is **open in Y** if V is an open set in the subspace topology on Y.*

Thus, a set is open in the subspace topology on Y if it is the intersection of an open set in X with Y.

Of course, we need to verify that the subspace topology is actually a topology:

(i) First, we note that $\varnothing$ and Y are both open in Y, since $\varnothing = \varnothing \cap Y$ and $Y = X \cap Y$.

(ii) Next, we show that intersections of finitely many open sets in Y are open in Y. Suppose that $V_1, \ldots, V_n$ are open in Y. Then for each i there exists U_i, open in X, such that $V_i = U_i \cap Y$. Now,

$$
\begin{aligned}
V_1 \cap \ldots \cap V_n &= (U_1 \cap Y) \cap \ldots \cap (U_n \cap Y) \\
&= (U_1 \cap \ldots \cap U_n) \cap Y.
\end{aligned}
$$

Since $U_1 \cap \ldots \cap U_n$ is open in X, it follows that $V_1 \cap \ldots \cap V_n$ is an open set in X intersected with Y, and therefore is open in Y.

(iii) Finally, we show that unions of arbitrarily many open sets in Y are open in Y. Thus, suppose that $\{V_\alpha\}$ is a collection of sets that are open in Y. Then for each α there exists an open set U_α in X such that $V_\alpha = U_\alpha \cap Y$. Therefore,

$$\bigcup V_\alpha = \bigcup (U_\alpha \cap Y) = \left(\bigcup U_\alpha \right) \cap Y.$$

Now, $\bigcup U_\alpha$ is open in X, implying that $\bigcup V_\alpha$ is an open set in X intersected with Y. Thus, $\bigcup V_\alpha$ is open in Y.

EXAMPLE 3.1. Consider $I = [0, 1]$ as a subset of $\mathbb{R}$ with the standard topology. In the subspace topology on I, open sets in I are open sets in $\mathbb{R}$ intersected with I. For example, sets of the form (a, b), where $0 < a < b < 1$, are open in the subspace topology on I (and open in $\mathbb{R}$ as well). Also, for $0 < a < 1$, the sets $[0, a)$ and $(a, 1]$ are open in the subspace topology on I, even though they are not open in $\mathbb{R}$. (See Figure 3.1.)
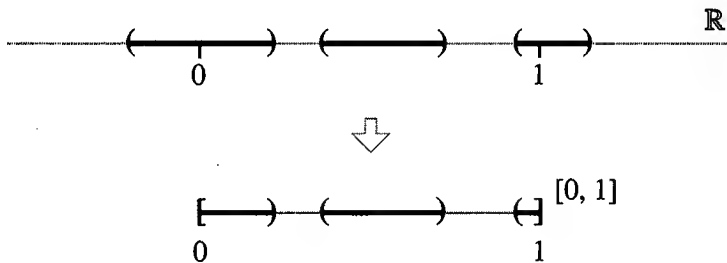
FIGURE 3.1: The subspace topology on $I = [0, 1]$.

EXAMPLE 3.2. Consider the subspace topology that the integers $\mathbb{Z}$ inherit from the standard topology on $\mathbb{R}$. (See Figure 3.2.) Since open intervals in $\mathbb{R}$ are open in the standard topology, and since each integer is contained in an open interval that contains no other integer, the single-point sets containing each integer are open sets in the subspace topology on $\mathbb{Z}$. But then arbitrary unions of these sets are open. Hence, every subset of $\mathbb{Z}$ is open in $\mathbb{Z}$, and the subspace topology on $\mathbb{Z}$ is the discrete topology.
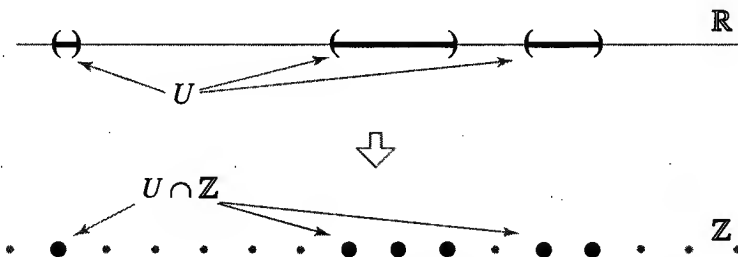
FIGURE 3.2: The topology that $\mathbb{Z}$ inherits from $\mathbb{R}$.

In the previous two examples, we can view the subspace topologies inherited from the standard topology on $\mathbb{R}$ as the standard topology on the corresponding sets. More generally, we have the following definition:

DEFINITION 3.2. *Let Y be a subset of $\mathbb{R}^n$. The **standard topology** on Y is the topology that Y inherits as a subspace of $\mathbb{R}^n$ with the standard topology.*

Thus the circle, the sphere, and, in general, the n-sphere all have a standard topology as subspaces of a Euclidean space. The same holds true for the disk, the n-ball, the open disk, and the open n-ball.

EXAMPLE 3.3. Let $Y = [-1, 0) \cup (0, 1] \subset \mathbb{R}$. In the standard topology on Y, both $[-1, 0)$ and $(0, 1]$ are open sets. Hence, their complements, which are $(0, 1]$ and $[-1, 0)$, respectively, are closed sets in the standard topology on Y. Thus, in the standard topology on Y, the sets $[-1, 0)$ and $(0, 1]$ are both open and closed.

In Definition 3.1 we defined "open in Y." Next we discuss closed sets in a subspace topology.

DEFINITION 3.3. *Let X be a topological space, and let $Y \subset X$ have the subspace topology. We say that a set $C \subset Y$ is **closed in** Y if C is closed in the subspace topology on Y.*

What does it mean for C to be closed in Y? As always, closed means that the complement is open, so a set $C \subset Y$ is closed in Y if $Y - C$ is open in Y. On the other hand, the following theorem indicates that we can obtain closed sets in Y by taking the intersection of closed sets in X with Y:

THEOREM 3.4. *Let X be a topological space, and let $Y \subset X$ have the subspace topology. Then $C \subset Y$ is closed in Y if and only if $C = D \cap Y$ for some closed set D in X.*

Proof. See Exercise 3.3. ∎

We will find it useful to be able to produce a basis for the subspace topology on a subset of a topological space, using a basis for the original space.

THEOREM 3.5. *Let X be a topological space and $\mathcal{B}$ be a basis for the topology on X. If $Y \subset X$, then the collection*

$$\mathcal{B}_Y = \{B \cap Y \mid B \in \mathcal{B}\}$$

is a basis for the subspace topology on Y.

Proof. First, note that $\mathcal{B}_Y$ is a collection of open sets in the subspace topology on Y. We use Theorem 1.13 to prove that $\mathcal{B}_Y$ is a basis for the subspace topology. Suppose W is an open set in the subspace topology on Y. Let $y \in W$ be arbitrary. Then $W = U \cap Y$, where U is open in X. There exists a basis element B in $\mathcal{B}$ such that $y \in B \subset U$. (See Figure 3.3.) Thus $y \in B \cap Y \subset U \cap Y = W$. Since $B \cap Y \in \mathcal{B}_Y$, it now follows by Theorem 1.13 that $\mathcal{B}_Y$ is a basis for the subspace topology on Y. ∎
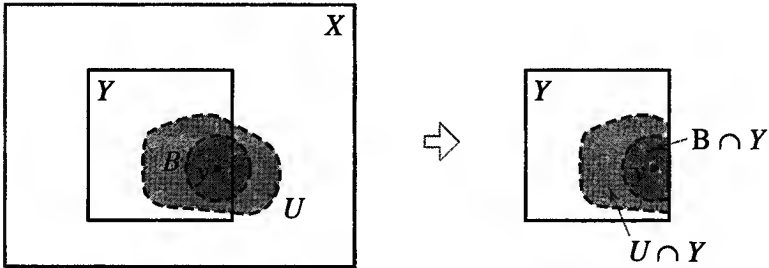
FIGURE 3.3: Inheriting basis elements for the subspace topology.

EXAMPLE 3.4. Consider the circle $S^1 \subset \mathbb{R}^2$ with the standard topology. Because open balls form a basis for the standard topology on $\mathbb{R}^2$, their intersection with S^1 forms a basis for the standard topology on S^1. The resulting basis elements are open intervals in the circle, consisting of all points between two angles in the circle. (See Figure 3.4.)
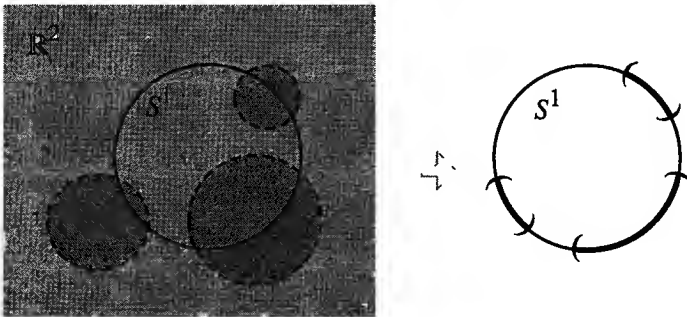
FIGURE 3.4: Basis elements for S^1 are open intervals in the circle.

IMPORTANT NOTE: *When we refer to "the circle" as a topological space, we mean S^1 with the standard topology. When we say "a circle" we mean a space topologically equivalent to the circle. The same holds for "the disk" versus "a disk," "the sphere" versus "a sphere," and so on.*

EXAMPLE 3.5. In $\mathbb{R}^3$, let C be the circle of radius 1 in the xz-plane with center at the point $(2, 0, 0)$. Consider the subspace of $\mathbb{R}^3$ swept out as C is rotated about the z-axis. (See Figure 3.5.) The resulting space is called the **torus** and is denoted by T.
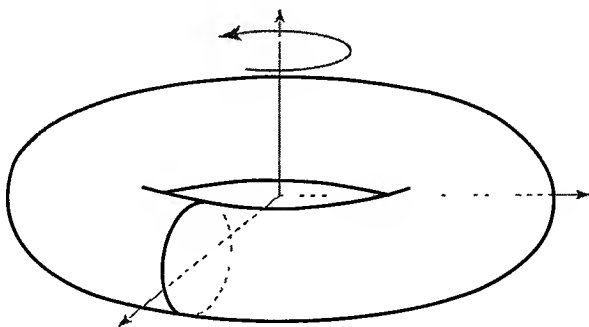
FIGURE 3.5: A circle rotated about the z-axis sweeps out the torus.

EXAMPLE 3.6. If S is a surface in $\mathbb{R}^3$, then the collection of open patches in S, obtained by intersecting open balls in $\mathbb{R}^3$ with S, is a basis for the standard topology on S. (See Figure 3.6.)
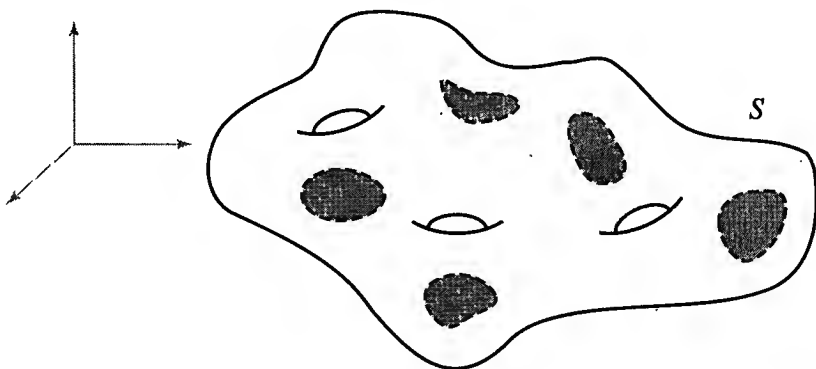
FIGURE 3.6: Open patches on the surface are basis elements for the standard topology.

Exercises for Section 3.1

3.1. Let $X = \{(x, 0) \in \mathbb{R}^2 \mid x \in \mathbb{R}\}$, the x-axis in the plane. Describe the topology that X inherits as a subspace of $\mathbb{R}^2$ with the standard topology.

3.2. Let $Y = [-1, 1]$ have the standard topology. Which of the following sets are open in Y and which are open in $\mathbb{R}$?

$$A = (-1, -1/2) \cup (1/2, 1)$$
$$B = (-1, -1/2] \cup [1/2, 1)$$
$$C = [-1, -1/2) \cup (1/2, 1]$$
$$D = [-1, -1/2] \cup [1/2, 1]$$
$$E = \bigcup_{n=1}^{\infty}(\tfrac{1}{1+n}, \tfrac{1}{n})$$

3.3. Prove Theorem 3.4: Let X be a topological space, and let $Y \subset X$ have the subspace topology. Then $C \subset Y$ is closed in Y if and only if $C = D \cap Y$ for some closed set D in X.

3.4. Let $Y = (0, 5]$. Which of the following subsets of Y are open, and which are closed, in Y in the standard topology?

    (a) $(0,1)$ (b) $(0,1]$ (c) $\{1\}$ (d) $(0,5]$ (e) $(1,2)$
    (f) $[1,2)$ (g) $(1,2]$ (h) $[1,2]$ (i) $(4,5]$ (j) $[4,5]$

3.5. Let $Y = (0, 5]$ have the subspace topology inherited from $\mathbb{R}$ with the lower limit topology. Which of the following subsets of Y are open, and which are closed, in Y in this topology?

    (a) $(0,1)$ (b) $(0,1]$ (c) $\{1\}$ (d) $(0,5]$ (e) $(1,2)$
    (f) $[1,2)$ (g) $(1,2]$ (h) $[1,2]$ (i) $(4,5]$ (j) $[4,5]$

3.6. Let $Y = (0, 4] \cup \{5\}$. Which of the following subsets of Y are open, and which are closed, in Y in the standard topology?

    (a) $(0,1)$ (b) $(0,1]$ (c) $\{1\}$ (d) $(0,4]$ (e) $(1,2)$
    (f) $[1,4)$ (g) $(1,4]$ (h) $[1,4]$ (i) $\{4\}$ (j) $\{4,5\}$

3.7. Let X be a Hausdorff topological space, and Y be a subset of X. Prove that the subspace topology on Y is Hausdorff.

3.8. Let X be a topological space, and let $Y \subset X$ have the subspace topology.
    (a) If A is open in Y, and Y is open in X, show that A is open in X.
    (b) If A is closed in Y, and Y is closed in X, show that A is closed in X.

3.9. (a) Let $K = \{\tfrac{1}{n} \in \mathbb{R} \mid n \in \mathbb{Z}_+\}$. Show that the standard topology on K is the discrete topology.
    (b) Let $K^* = K \cup \{0\}$. Show that the standard topology on K^* is not the discrete topology.

3.10. Show that the standard topology on $\mathbb{Q}$, the set of rational numbers, is not the discrete topology.

3.11. Let A be a subspace of X, and $D \subset A$. By $\mathrm{Int}_A D$ and $\mathrm{Int}_X D$ we mean, respectively, the interior of D in the subspace topology on A and the interior of D in the topology on X. Similarly, we define $\mathrm{Cl}_A D$ and $\mathrm{Cl}_X D$ for the closure, and we define $\partial_A D$ and $\partial_X D$ for the boundary.
    (a) Show that $\mathrm{Int}_A D \subset A \cap \mathrm{Int}_X D$, and provide an example showing that equality does not hold in general.
    (b) Explore the relationship between $\mathrm{Cl}_A D$ and $A \cap \mathrm{Cl}_X D$. For each containment $\subset$ and $\supset$, either prove that it holds or find a counterexample.
    (c) Explore the relationship between $\partial_A D$ and $A \cap \partial_X D$. For each containment $\subset$ and $\supset$, either prove that it holds or find a counterexample.

3.2 The Product Topology

Given two topological spaces X and Y, we would like to generate a natural topology on the product, $X \times Y$. Our first inclination might be to take as the topology on $X \times Y$ the collection $\mathcal{C}$ of sets of the form $U \times V$ where U is open in X and V is open in Y. But $\mathcal{C}$ is not a topology since the union of two sets $U_1 \times V_1$ and $U_2 \times V_2$ need not be in the form $U \times V$ for some $U \subset X$ and $V \subset Y$. (See Figure 3.7.) However, if we use $\mathcal{C}$ as a basis, rather than as the whole topology, we can proceed.
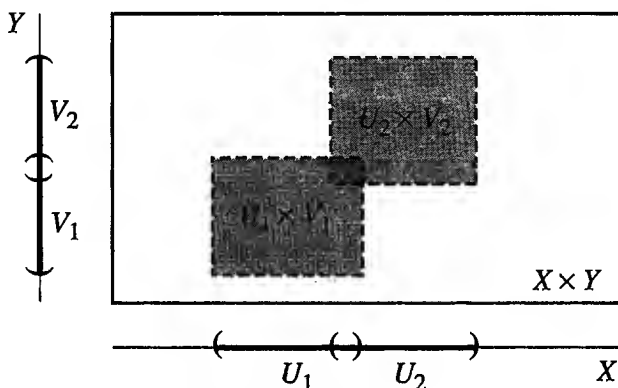
FIGURE 3.7: The union of the sets $U_1 \times V_1$ and $U_2 \times V_2$ need not be in the form $U \times V$.

DEFINITION 3.6. *Let X and Y be topological spaces and $X \times Y$ be their product. The* **product topology** *on $X \times Y$ is the topology generated by the basis*

$$\mathcal{B} = \{U \times V \mid U \text{ is open in } X \text{ and } V \text{ is open in } Y\}.$$

Of course, we must verify that $\mathcal{B}$ actually is a basis for a topology on the product, $X \times Y$.

THEOREM 3.7. *The collection $\mathcal{B}$ is a basis for a topology on $X \times Y$.*

Proof. Every point (x, y) is in $X \times Y$, and $X \times Y \in \mathcal{B}$. Therefore the first condition for a basis is satisfied.

Next assume that (x, y) is in the intersection of two elements of $\mathcal{B}$. That is, $(x, y) \in (U_1 \times V_1) \cap (U_2 \times V_2)$ where U_1 and U_2 are open sets in X, and V_1 and V_2 are open sets in Y. Let $U_3 = U_1 \cap U_2$ and $V_3 = V_1 \cap V_2$. Then U_3 is open in X, and V_3 is open in Y, and therefore $U_3 \times V_3 \in \mathcal{B}$. Also,

$$U_3 \times V_3 = (U_1 \cap U_2) \times (V_1 \cap V_2) = (U_1 \times V_1) \cap (U_2 \times V_2),$$

and thus $(x, y) \in U_3 \times V_3 \subset (U_1 \times V_1) \cap (U_2 \times V_2)$. It follows that the second condition for a basis is satisfied.

Therefore $\mathcal{B}$ is a basis for a topology on $X \times Y$.   ∎

EXAMPLE 3.7. Let $X = \{a, b, c\}$ and $Y = \{1, 2\}$ with topologies $\{\varnothing, \{b\}, \{c\}, \{a, b\}, \{b, c\}, X\}$ and $\{\varnothing, \{1\}, Y\}$, respectively. A basis for the product topology on $X \times Y$ is depicted in Figure 3.8. Each nonempty open set in the product topology on $X \times Y$ is a union of the basis elements shown.
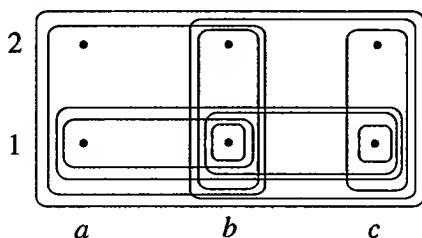
FIGURE 3.8: A basis for the product topology on $X \times Y$.

IMPORTANT NOTE: *As we indicated at the outset of this section, it is not only the products of open sets that are open sets in the product topology. Those sets, along with all possible unions of them, are what make up the product topology.*

As with open sets, products of closed sets are closed sets in the product topology. (See Exercise 3.19.) But here too, this does not account for all of the closed sets because there are closed sets in the product topology that cannot be expressed as a product of closed sets. For instance, the set $\{(a, 2), (c, 1), (c, 2)\}$ is a closed set in the product topology in Example 3.7, but it is not a product of closed sets.

In Definition 3.6, the basis $\mathcal{B}$ that we use to define the product topology is relatively large since we obtain it by pairing up every open set U in X with every open set V in Y. Fortunately, as the next theorem indicates, we can find a smaller basis for the product topology by using bases for the topologies on X and Y, rather than using the whole topologies themselves.

THEOREM 3.8. *If C is a basis for X and $\mathcal{D}$ is a basis for Y, then*

$$\mathcal{E} = \{C \times D \mid C \in \mathcal{C} \text{ and } D \in \mathcal{D}\}$$

is a basis that generates the product topology on $X \times Y$.

> **Proof.** Each set $C \times D \in \mathcal{E}$ is an open set in the product topology; therefore, by Theorem 1.13, it suffices to show that for every open set W in $X \times Y$ and every point $(x, y) \in W$, there is a set $C \times D$ in $\mathcal{E}$ such that $(x, y) \in C \times D \subset W$. But since W is open in $X \times Y$, we know that there are open sets U in X and V in Y such that $(x, y) \in U \times V \subset W$. So $x \in U$ and $y \in V$. Since U is open in X, there is a basis element $C \in \mathcal{C}$ such that $x \in C \subset U$. Similarly, since V is open in Y, there is a basis element $D \in \mathcal{D}$ such that $y \in D \subset V$. Thus $(x, y) \in C \times D \subset U \times V \subset W$. Hence, by Theorem 1.13, it follows that $\mathcal{E} = \{C \times D \mid C \in \mathcal{C} \text{ and } D \in \mathcal{D}\}$ is a basis for the product topology on $X \times Y$. ∎

EXAMPLE 3.8. Let $I = [0, 1]$ have the standard topology as a subspace of $\mathbb{R}$. The product space $I \times I$ is called the **unit square**. (See Figure 3.9.) The product topology on $I \times I$ is the same as the standard topology on $I \times I$ as a subspace of $\mathbb{R}^2$. (See Theorem 3.9.)
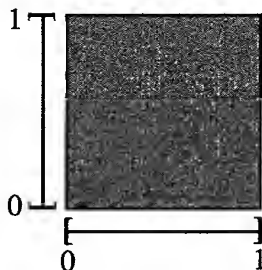
FIGURE 3.9: The unit square, $I \times I$.

IMPORTANT NOTE: *When picturing a product space, we can imagine a copy of the second space corresponding to each point in the first space, or vice versa.*

EXAMPLE 3.9. Let S^1 be the circle, and let $I = [0, 1]$ have the standard topology. Then $S^1 \times I$ appears as in Figure 3.10. We can think of it as a circle with intervals perpendicular at each point of the circle. Seen this way, it is a circle's worth of intervals. Or it can be thought of as an interval with perpendicular circles at each point. Thus it is an interval's worth of circles. The resulting topological space is called the **annulus**.

The product space $S^1 \times (0, 1)$ is the annulus with the innermost and outermost circles removed. We refer to it as the **open annulus**.
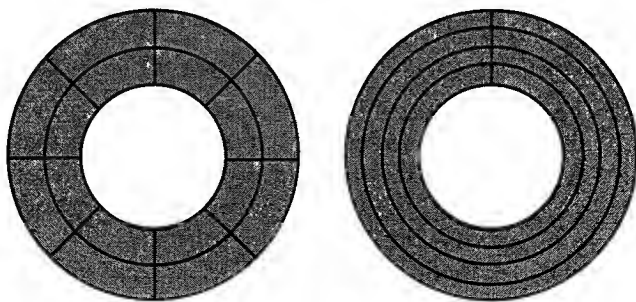
FIGURE 3.10: Depicting $S^1 \times I$.

EXAMPLE 3.10. Consider the product space $S^1 \times S^1$, where S^1 is the circle. For each point in the first S^1, there is a circle corresponding to the second S^1. (See Figure 3.11.) Since each S^1 has a topology generated by open intervals in the circle, it follows by Theorem 3.8 that $S^1 \times S^1$ has a basis consisting of rectangular open patches, as shown in the figure. The resulting space resembles the torus introduced in Example 3.5; in fact, they are topologically equivalent.
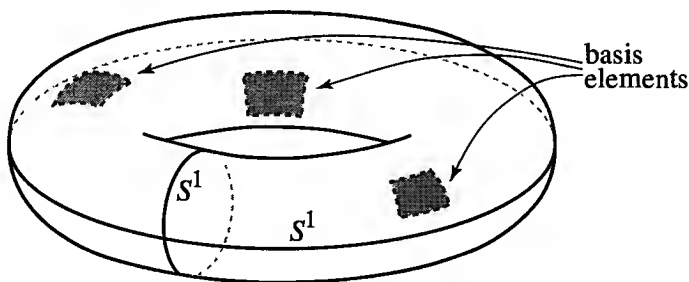
FIGURE 3.11: The product $S^1 \times S^1$ is a torus.

EXAMPLE 3.11. Let D be the disk as a subspace of the plane. The product space $S^1 \times D$ is called the **solid torus**. (See Figure 3.12.) If we think of the torus as the surface of a doughnut, then the solid torus is the whole doughnut itself.
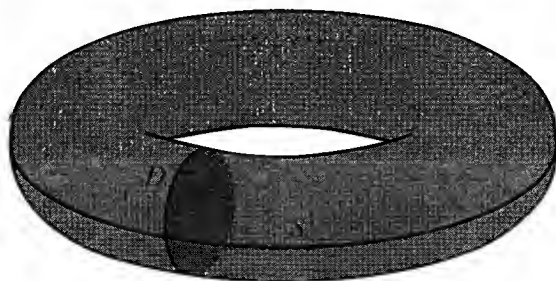
FIGURE 3.12: The product $S^1 \times D$ is the solid torus.

Let A and B be subsets of topological spaces X and Y, respectively. We now have two natural ways to put a topology on $A \times B$. On the one hand, we can view $A \times B$ as a subspace of the product $X \times Y$. On the other hand, we can view $A \times B$ as the product of subspaces, $A \subset X$ and $B \subset Y$. The next theorem indicates that both approaches result in the same topology.

THEOREM 3.9. *Let X and Y be topological spaces, and assume that $A \subset X$ and $B \subset Y$. Then the topology on $A \times B$ as a subspace of the product $X \times Y$ is the same as the product topology on $A \times B$, where A has the subspace topology inherited from X, and B has the subspace topology inherited from Y.*

   Proof. See Exercise 3.15.   ■

   The approach used to define a product of two spaces extends to a product $X_1 \times \ldots \times X_n$ of n topological spaces. It is straightforward to see that the collection

$$\mathcal{B} = \{U_1 \times \ldots \times U_n \mid U_i \text{ is open in } X_i \text{ for each } i\}$$

is a basis for a topology on $X_1 \times \ldots \times X_n$. The resulting topology is called the **product topology** on $X_1 \times \ldots \times X_n$. We have an analog to Theorem 3.8 for this case. Specifically, if $\mathcal{B}_i$ is a basis for X_i for each $i = 1, \ldots, n$, then the collection

$$\mathcal{B}' = \{B_1 \times \ldots \times B_n \mid B_i \in \mathcal{B}_i \text{ for } i = 1, \ldots, n\}$$

is a basis for $X_1 \times \ldots \times X_n$.

   Given topological spaces X, Y, and Z, the space $(X \times Y) \times Z$ is the product of the spaces $X \times Y$ and Z, the space $X \times (Y \times Z)$ is the product of the spaces X and $Y \times Z$, and the space $X \times Y \times Z$ is the product of the three spaces X, Y, and Z together, as described above. These three spaces are naturally topologically equivalent. (See Exercise 4.34.) Therefore, we do not distinguish between these spaces, and we always use the latter notation, without grouping, to denote a product of multiple spaces.

   In Section 1.2 we defined the standard topology on $\mathbb{R}^n$ to be the topology generated by the basis of open balls defined by the Euclidean distance formula on $\mathbb{R}^n$. We also pointed out in Section 1.2 that the same topology results from taking a basis made up of products of open intervals in $\mathbb{R}$. It follows that the standard topology on $\mathbb{R}^n$ is the same as the product topology that results from taking the product of n copies of $\mathbb{R}$ with the standard topology.

EXAMPLE 3.12. The n-**torus**, T^n, is the topological space obtained by taking the product of n copies of the circle, S^1.

   The next theorem indicates that, nicely enough, the interior of a product is the product of the interiors.

THEOREM 3.10. *Let A and B be subsets of topological spaces X and Y, respectively. Then $Int(A \times B) = Int(A) \times Int(B)$.*

   Note that in the statement of the theorem, $Int(\ )$ is taking the interior in three separate topological spaces—$Int(A \times B)$ is the interior in $X \times Y$, $Int(A)$ is the interior in X, and $Int(B)$ is the interior in Y.

Proof. Since Int(A) is an open set contained in A, and Int(B) is an open set contained in B, it follows that Int(A) $\times$ Int(B) is an open set in the product topology and is contained in $A \times B$. Thus Int(A) $\times$ Int(B) $\subset$ Int($A \times B$).

Now suppose $(x, y) \in$ Int($A \times B$). We prove that $(x, y) \in$ Int(A) $\times$ Int(B). Since $(x, y) \in$ Int($A \times B$), it follows that (x, y) is contained in an open set contained in $A \times B$ and therefore is also contained in a basis element contained in $A \times B$. So there exists a U and V, open in X and Y, respectively, such that $(x, y) \in U \times V \subset A \times B$. Thus, x is in an open set U contained in A, and y is in an open set V contained in B, implying that $x \in$ Int(A) and $y \in$ Int(B). Therefore $(x, y) \in$ Int(A) $\times$ Int(B). It follows that Int($A \times B$) $\subset$ Int(A) $\times$ Int(B).

Since we have both Int(A) $\times$ Int(B) $\subset$ Int($A \times B$) and Int($A \times B$) $\subset$ Int(A) $\times$ Int(B), the desired equality follows.   ∎

Results parallel to Theorem 3.10, for the closure and boundary of a product of sets, are considered in Exercises 3.20 and 3.22.

Exercises for Section 3.2

3.12. Is the finite complement topology on $\mathbb{R}^2$ the same as the product topology on $\mathbb{R}^2$ that results from taking the product $\mathbb{R}_{fc} \times \mathbb{R}_{fc}$, where $\mathbb{R}_{fc}$ is $\mathbb{R}$ in the finite complement topology? Justify your answer.

3.13. Let $X = PP\mathbb{R}^2_{(0,0)}$, the particular point topology on $\mathbb{R}^2$ with the origin serving as the particular point. (See Exercise 1.7.) Is X the same as the topology that results from taking the product of $\mathbb{R}$ with itself, where each $\mathbb{R}$ has the particular point topology $PP\mathbb{R}_0$? Justify your answer.

3.14. Prove that the digital plane introduced in Section 1.4 is the topological space that results from taking the product of two copies of the digital line.

3.15. Prove Theorem 3.9: Let X and Y be topological spaces, and assume that $A \subset X$ and $B \subset Y$. Then the topology on $A \times B$ as a subspace of the product $X \times Y$ is the same as the product topology on $A \times B$, where A has the subspace topology inherited from X, and B has the subspace topology inherited from Y.

3.16. Let S^2 be the sphere, D be the disk, T be the torus, S^1 be the circle, and $I = [0, 1]$ with the standard topology. Draw pictures of the product spaces $S^2 \times I, T \times I, S^1 \times I \times I$, and $S^1 \times D$.

3.17. If L is a line in the plane, describe the subspace topology it inherits from $\mathbb{R}_l \times \mathbb{R}$ and from $\mathbb{R}_l \times \mathbb{R}_l$, where $\mathbb{R}_l$ is the real line in the lower limit topology. Note that the result depends on the slope of the line. In all cases, it is a familiar topology.

3.18. Show that if X and Y are Hausdorff spaces, then so is the product space $X \times Y$.

3.19. Show that if A is closed in X and B is closed in Y, then $A \times B$ is closed in $X \times Y$.

3.20. Show that if $A \subset X$ and $B \subset Y$, then Cl($A \times B$) $=$ Cl(A) $\times$ Cl(B).

3.21. Determine whether or not the sets in Figure 3.13 are open, closed, both, or neither in the product topologies on the plane given by $\mathbb{R} \times \mathbb{R}$, $\mathbb{R}_l \times \mathbb{R}$, and $\mathbb{R}_l \times \mathbb{R}_l$, where $\mathbb{R}_l$ is the real line in the lower limit topology.
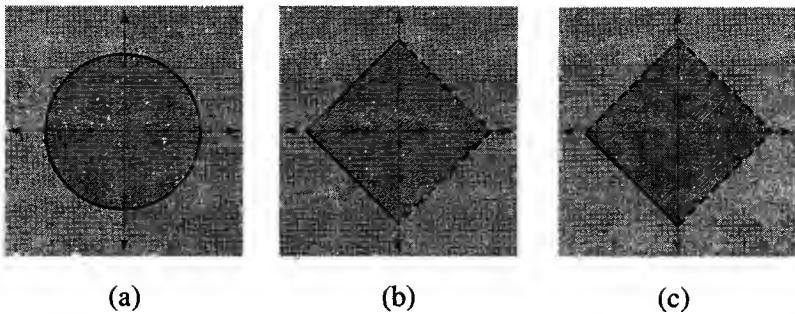
FIGURE 3.13: Are these sets open, closed, both, or neither in $\mathbb{R} \times \mathbb{R}$, $\mathbb{R}_l \times \mathbb{R}$, and $\mathbb{R}_l \times \mathbb{R}_l$?

3.22. Suppose that $A \subset X$ and $B \subset Y$.
   (a) Provide an example demonstrating that $\partial(A \times B) = \partial(A) \times \partial(B)$ does not hold in general.
   (b) Derive and prove a relationship expressing $\partial(A \times B)$ in terms of $\partial(A)$, $\partial(B)$, A, and B.

3.3 The Quotient Topology

The concept of a quotient topology allows us to construct a variety of additional topological spaces from the ones that we have already introduced. Put simply, we create a topological model that mimics the process of gluing together or collapsing parts of one or more objects. One of the most well-known examples is the torus, as obtained from a square sheet by gluing together the opposite edges. (See Figure 3.14.)
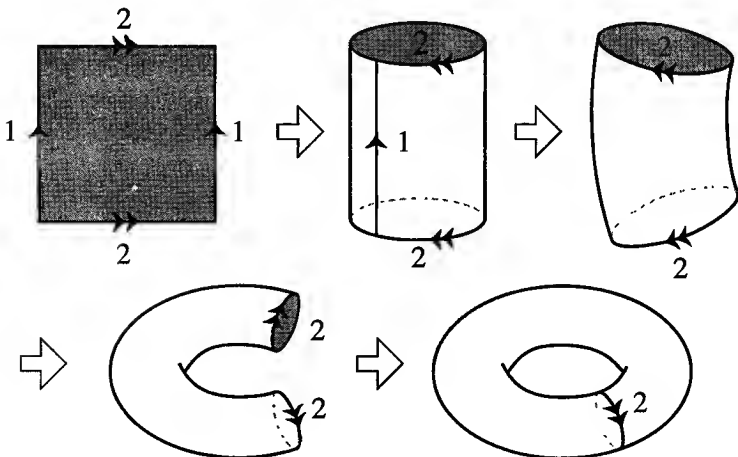
FIGURE 3.14: Gluing together opposite edges of a square to obtain the torus.

The quotient topology is defined as follows:

DEFINITION 3.11. *Let X be a topological space and A be a set (that is not necessarily a subset of X). Let p : X → A be a surjective map. Define a subset U of A to be open in A if and only if $p^{-1}(U)$ is open in X. The resultant collection of open sets in A is called the **quotient topology induced by** p , and the function p is called a **quotient map**. The topological space A is called a quotient space.*

It might not be clear how this definition of a topology enables us to glue the opposite edges of a square to obtain a torus, but we will see this in detail in Example 3.20. First, we verify that the quotient topology is a topology:

THEOREM 3.12. *Let p : X → A be a quotient map. The quotient topology on A induced by p is a topology.*

Proof. We verify each of the three conditions for a topology.

(i) The set $p^{-1}(\varnothing) = \varnothing$, which is open in X. The set $p^{-1}(A) = X$, which is open in X. So $\varnothing$ and A are open in the quotient topology.

(ii) Suppose each of the sets in the collection $\{U_\beta\}_{\beta \in B}$ is open in the quotient topology on A. Then $p^{-1}(\bigcup(U_\beta)) = \bigcup p^{-1}(U_\beta)$, which is a union of open sets in X, and therefore is open in X. Thus, $\bigcup(U_\beta)$ is open in the quotient topology, implying that the arbitrary union of open sets in the quotient topology is an open set in the quotient topology.

(iii) Suppose each of the sets U_i, $i = 1, \ldots, n$, is open in the quotient topology on A. Then $p^{-1}(\bigcap U_i) = \bigcap p^{-1}(U_i)$, which is a finite intersection of open sets in X, and therefore is open in X. Hence, $\bigcap U_i$ is open in the quotient topology, and it follows that the finite intersection of open sets in the quotient topology is an open set in the quotient topology.

Hence, the quotient topology is a topology on A.   ∎

EXAMPLE 3.13. Give $\mathbb{R}$ the standard topology, and define

$$p : \mathbb{R} \to \{a, b, c\} \text{ by } p(x) = \begin{cases} a \text{ if } x < 0, \\ b \text{ if } x = 0, \\ c \text{ if } x > 0. \end{cases}$$

The resulting quotient topology on $\{a, b, c\}$ is depicted in Figure 3.15. The subsets $\{a\}$, $\{c\}$, and $\{a, c\}$ are all open since their preimages are open in $\mathbb{R}$. But $\{b\}$ is not open since its preimage is $\{0\}$, which is not open in $\mathbb{R}$.
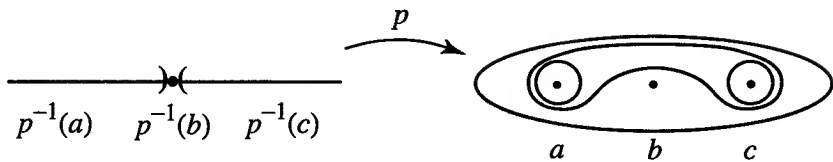
FIGURE 3.15: The quotient topology on $\{a, b, c\}$ induced by the function p.

EXAMPLE 3.14. Let $\mathbb{R}$ have the standard topology, and define $p : \mathbb{R} \to \mathbb{Z}$ by $p(x) = x$ if x is an integer, and $p(x) = n$ if $x \in (n - 1, n + 1)$ and n is an odd integer. (See Figure 3.16.) So p is the identity on the integers, and p maps noninteger values to the nearest odd integer.
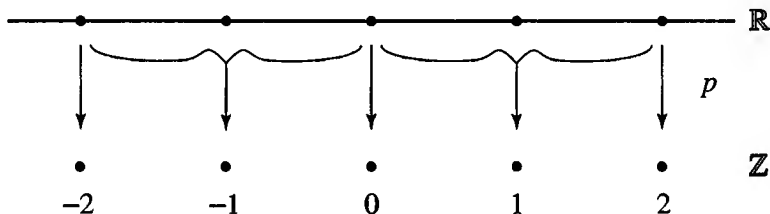
FIGURE 3.16: The function $p : \mathbb{R} \to \mathbb{Z}$.

In the resulting quotient topology on $\mathbb{Z}$, if n is an odd integer, then $\{n\}$ is an open set since $p^{-1}(\{n\}) = (n - 1, n + 1)$, an open set in $\mathbb{R}$. If n is an even integer, then $\{n\}$ is not an open set since $p^{-1}(\{n\})$ is not open in $\mathbb{R}$. In the quotient topology, the smallest open set containing an even integer n is the set $\{n - 1, n, n + 1\}$.

It follows that the quotient topology induced by p on $\mathbb{Z}$ is the digital line topology introduced in Example 1.10. This quotient mapping demonstrates a natural passage from the line to a discrete model of it, an idea that is central in digital image processing. (See Sections 1.4 and 11.3.) The line is subdivided into open pixels—the intervals $(n - 1, n + 1)$, for n odd—along with the boundaries between the pixels—the points n, for n even. Then, in the quotient space the pixels become the single-point open sets $\{n\}$, for n odd, and the boundaries between the pixels become the single-point closed sets $\{n\}$, for n even.

Let X be a topological space. We are particularly interested in quotient spaces defined on partitions of X. Specifically, let X^* be a collection of mutually disjoint subsets of X whose union is X, and let $p : X \to X^*$ be the surjective map that takes each point in X to the corresponding element of X^* that contains it. Then p induces a quotient topology on X^*. We think of the process of going from the topology on X to the quotient topology on X^* as taking each subset S in the partition and identifying all of the points in S with

one another, thereby collapsing S to a single point in the quotient space. (See Figure 3.17.) A set U of points in X^* is open in the quotient topology on X^* exactly when the union of the subsets of X, corresponding to the points in U, is an open subset in X.
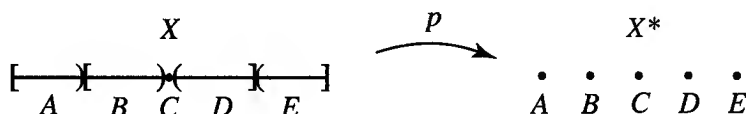
FIGURE 3.17: Collapsing each set in the partition to a point in the quotient space.

EXAMPLE 3.15. In Figure 3.18 we illustrate a quotient space defined on a partition. Let $X = \{a, b, c, d, e\}$ with topology

$$\{\varnothing, \{a\}, \{a, b\}, \{a, b, c\}, \{a, b, c, d\}, X\}.$$

With $A = \{a, b\}$ and $B = \{c, d, e\}$, let X^* be the partition of X given by $X^* = \{A, B\}$. Note that X^* is a two-point set. Since $\{a, b\}$ is open in X and $\{c, d, e\}$ is not, the only open sets in the quotient topology on X^* are $\varnothing$, $\{A\}$, and X^* itself.
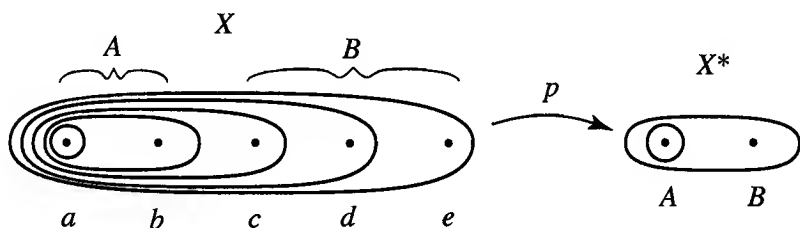
FIGURE 3.18: The space X and the quotient space determined by the partition $X^* = \{A, B\}$.

EXAMPLE 3.16. Let $X = [0, 1]$, and consider the partition X^* that is made up of the single-point sets $\{x\}$, for $0 < x < 1$, and the double-point set $D = \{0, 1\}$. (See Figure 3.19.) Then, in the quotient topology on X^*, we think of D as a single point, as if we had glued the two endpoints of $[0, 1]$ together.

A subset of X^* that does not contain D is a collection of single-point subsets, and it is open in X^* exactly when the union of those single-point sets is an open subset of $(0, 1)$. A subset of X^* that contains D is open in X^* when the union of all the sets making up the subset is an open subset of $[0, 1]$. Such

an open subset must contain 0 and 1, and therefore must contain intervals $[0, a)$ and $(b, 1]$, which are open in the subspace topology on $[0, 1]$. The resulting space is topologically equivalent to the circle, S^1.
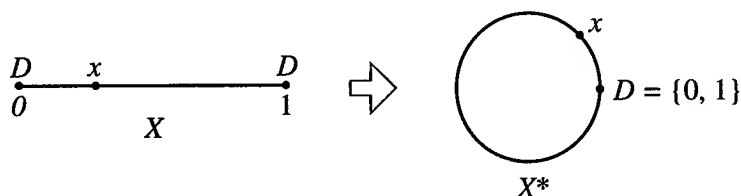
FIGURE 3.19: Forming a quotient topology on $[0, 1]$ by gluing the point 0 to the point 1.

EXAMPLE 3.17. In the previous example, we glued the endpoints of an interval together to obtain a single point. That is an example of a more general construction that results in a space known as a topological graph. Specifically, a **topological graph** G is a quotient space constructed by taking a finite set of points, called the **vertices** of G, along with a finite set of mutually disjoint closed bounded intervals in $\mathbb{R}$, and gluing the endpoints of the intervals to the vertices in some fashion. (See Figure 3.20.) The glued intervals are called the **edges** of G.

We examine topological graphs and their role in graph theory further in Chapter 13.
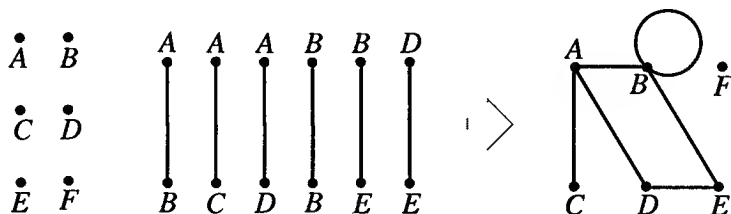
FIGURE 3.20: A topological graph is obtained by gluing endpoints of intervals to a set of vertices.

EXAMPLE 3.18. In Example 3.16 we obtained a circle by identifying endpoints of an interval in the real line. We describe a similar process here, using the digital line, that yields spaces we call digital circles.

Specifically, a **digital interval** is a subset $\{m, m + 1, \ldots, n\}$ of $\mathbb{Z}$ with the subspace topology inherited from the digital line topology. Let I_n be the digital interval in the form $\{1, 2, \ldots, n - 1, n\}$. If $n \geq 5$ is an odd integer,

then the topological space C_{n-1} resulting from identifying the endpoints 1 and n in I_n is called a **digital circle**. In Figure 3.21 we illustrate I_7 and C_6 along with a basis for each. By definition, a digital circle contains an even number of points.
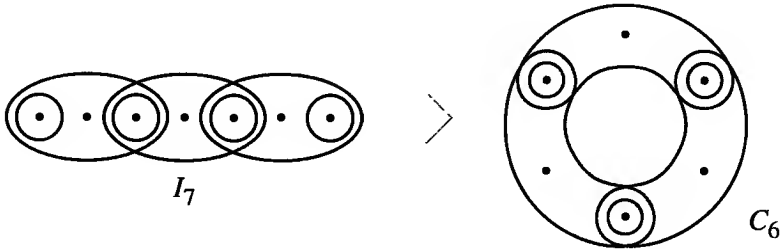
FIGURE 3.21: The digital circle C_6 is a quotient space of the digital interval I_7.

In Exercise 3.31 we examine the topological spaces that result when we take a general digital interval $\{m, m+1, \ldots, n\}$ and identify the endpoints m and n. Digital circles arise only when both m and n are odd.

EXAMPLE 3.19. Here we look at two different quotient spaces defined on $I \times I$. In the first case, we define a partition by taking subsets of the following form:

    (i) $A_{x,y} = \{(x, y)\}$ for every x and y such that $0 < x < 1$ and $0 \le y \le 1$.

    (ii) $B_y = \{(0, y), (1, y)\}$ for every y such that $0 < y < 1$.

In the quotient topology, the subsets B_y cause the left and right edges of the square to be glued as shown in Figure 3.22. The result is a space that is topologically equivalent to the annulus.
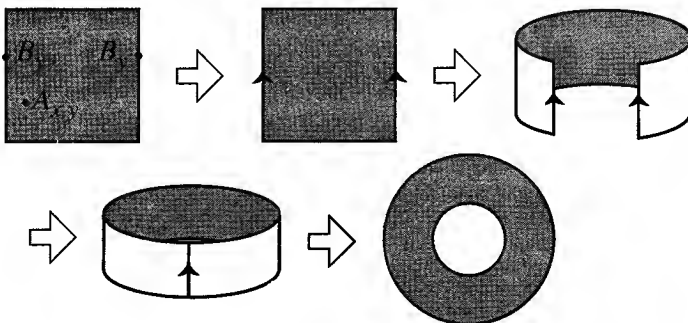
FIGURE 3.22: Obtaining the annulus by gluing the left and right edges of a square.

In the second case, we define a partition by taking subsets of the following form:

(i) $A_{x,y} = \{(x, y)\}$ for every x and y such that $0 < x < 1$ and $0 \le y \le 1$.

(ii) $B_y^* = \{(0, y), (1, 1 - y)\}$ for every y such that $0 < y < 1$.

Here the subsets B_y^* also cause the left and right edges of the square to be glued. But, as we see in Figure 3.23, in order to accomplish the gluing, we need to perform a half twist so that the identified points on the edges can be properly brought together. The result is the well-known **Möbius band**.
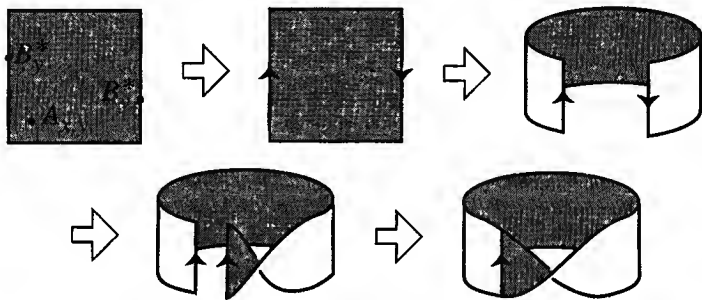
FIGURE 3.23: Obtaining the Möbius band by gluing the left and right edges of a square.

In situations similar to Example 3.19, we denote the gluings that we do by putting direction arrows on each of a pair of edges whose points are identified. Then we match the points on the two edges in a one-to-one fashion so that the directions on the glued edges coincide.

EXAMPLE 3.20. Define a partition of $I \times I$ by taking subsets of the following form:

(i) $A_{x,y} = \{(x, y)\}$ for every x and y such that $0 < x < 1$ and $0 < y < 1$.

(ii) $B_y = \{(0, y), (1, y)\}$ for every y such that $0 < y < 1$.

(iii) $C_x = \{(x, 0), (x, 1)\}$ for every x such that $0 < x < 1$.

(iv) $D = \{(0, 0), (0, 1), (1, 0), (1, 1)\}$.

In the quotient topology, the two-point subsets in (ii) cause the gluing of the left edge of the square to the right edge, and the two-point subsets in (iii) cause the gluing of the top edge of the square to the bottom edge. (See Figure 3.24.) Furthermore, the four-point subset causes the gluing of the four corners of the square to a single point. The topological space we obtain is therefore the result

of taking a square and gluing together its opposite edges. As shown previously in Figure 3.14, such a construction results in a torus.
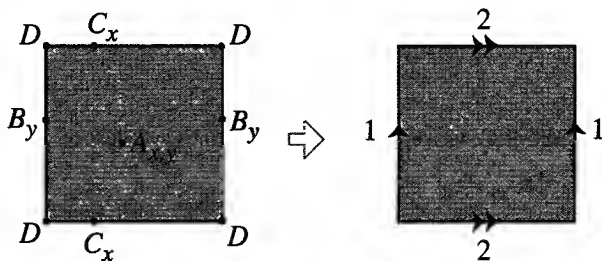
FIGURE 3.24: Partitioning the square to obtain a torus.

In Figure 3.24, even though we do not explicitly show that the four corners of the square are identified, note that the gluing of the edges labeled 1 and the gluing of the edges labeled 2 identify all four corners with each other.

An advantage to this representation of the torus is that we can imagine what it is like to move around on the torus sitting in 3-space, without ever leaving the flat two-dimensional representation that keeps track of the gluings. Each time we reach a point on the top edge of the square, we are transported to the corresponding point on the bottom edge. This process allows us to view a space in a setting of one dimension less than the setting in which it naturally resides. We will find this approach to viewing a space particularly useful when visualizing higher-dimensional spaces in Chapter 14.

This idea of a square with opposite edges identified occurs in some computer games. If we are moving around on the screen and we go off the left edge of the screen, we reappear at the same height on the right edge. If we go off the top edge, we reappear on the bottom edge. So, in fact, these games are actually being played on the torus.

In the next section, we examine more examples of spaces obtained by identifying points on the edges of polygons, and we examine a few three-dimensional examples obtained by gluing faces of a cube.

Exercises for Section 3.3

3.23. If $\mathbb{R}$ has the standard topology, define

$$p : \mathbb{R} \to \{a, b, c, d, e\} \text{ by } p(x) = \begin{cases} a \text{ if } x > 2, \\ b \text{ if } x = 2, \\ c \text{ if } 0 \leq x < 2, \\ d \text{ if } -1 < x < 0, \\ e \text{ if } x \leq -1. \end{cases}$$

(a) List the open sets in the quotient topology on $\{a, b, c, d, e\}$.

(b) Now assume that $\mathbb{R}$ has the lower limit topology. What are the open sets in the resulting quotient topology on $\{a, b, c, d, e\}$?

3.24. Let $X = \mathbb{R}$ in the standard topology. Take the partition

$$X^* = \{\ldots, (-1, 0], (0, 1], (1, 2], \ldots\}.$$

Describe the open sets in the resulting quotient topology on X^*.

3.25. Define a partition of $X = \mathbb{R}^2 - \{O\}$ by taking each ray emanating from the origin as an element in the partition. (See Figure 3.25.) Which topological space that we have previously encountered appears to be topologically equivalent to the quotient space that results from this partition?
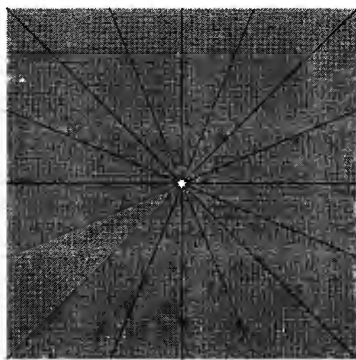
FIGURE 3.25: A partition of $X = \mathbb{R}^2 - \{O\}$.

3.26. Devise the rules for a game of tic-tac-toe that is played on the surface of the torus, using the square model of the torus as illustrated in Figure 3.26. What new three-in-a-rows would this game allow that are not allowed in the standard game?
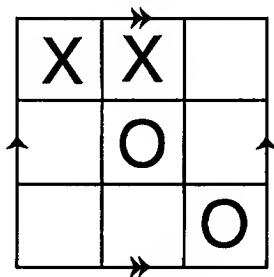
FIGURE 3.26: Tic-tac-toe on a torus.

3.27. Provide an example showing that a quotient space of a Hausdorff space need not be a Hausdorff space.

3.28. Consider the equivalence relation on $\mathbb{R}$ defined by $x \sim y$ if $x - y \in \mathbb{Z}$. Describe the quotient space that results from the partition of $\mathbb{R}$ into the equivalence classes in the equivalence relation.

3.29. Consider the equivalence relation on $\mathbb{R}^2$ defined by $(x_1, x_2) \sim (w_1, w_2)$ if $x_1 + x_2 = w_1 + w_2$. Describe the quotient space that results from the partition of $\mathbb{R}^2$ into the equivalence classes in this equivalence relation.

3.30. Consider the equivalence relation on $\mathbb{R}^2$ defined by $(x_1, x_2) \sim (w_1, w_2)$ if $x_1^2 + x_2^2 = w_1^2 + w_2^2$. Describe the quotient space that results from the partition of $\mathbb{R}^2$ into the equivalence classes in this equivalence relation.

3.31. Consider the four cases of m and n being either even or odd, and describe the topological spaces that result when we identify the endpoints m and n in a general digital interval $\{m, m + 1, \ldots, n\}$.

3.32. (a) Show that the digital line can be obtained as a quotient space that results from a partition of $\mathbb{R}$ in the standard topology.
   (b) Show that the digital plane, introduced in Section 1.4, can be obtained as a quotient space that results from a partition of $\mathbb{R}^2$ in the standard topology.

3.33. In each of the following cases, describe or draw a picture of the resulting quotient space. Assume that points are identified only with themselves unless they are explicitly said to be identified with other points.
   (a) The disk with its boundary points identified with each other to form a single point.
   (b) The circle S^1 with each pair of antipodal points identified with each other.
   (c) The interval $[0, 4]$, as a subspace of $\mathbb{R}$, with integer points identified with each other.
   (d) The interval $[0, 9]$, as a subspace of $\mathbb{R}$, with even integer points identified with each other to form a point and with odd integer points identified with each other to form a different point.
   (e) The real line $\mathbb{R}$ with $[-1, 1]$ collapsed to a point.
   (f) The real line $\mathbb{R}$ with $[-2, -1] \cup [1, 2]$ collapsed to a point.
   (g) The real line $\mathbb{R}$ with $(-1, 1)$ collapsed to a point.
   (h) The plane $\mathbb{R}^2$ with the circle S^1 collapsed to a point.
   (i) The plane $\mathbb{R}^2$ with the circle S^1 and the origin collapsed to a point.
   (j) The sphere with the north and south pole identified with each other.
   (k) The sphere with the equator collapsed to a point.

3.34. An optical character-recognition program attempts to recognize a character (for example, a letter of the alphabet) from a computer image of it. Topological properties of each letter's shape can help in this identification. Create a topological graph representation of each letter of the alphabet, and in each case use a topological graph with as few edges as possible.

3.4 More Examples of Quotient Spaces

We ended the previous section with a description of how to obtain a torus by gluing together opposite edges in a square. In this section, we examine other quotient spaces obtained by gluing together parts of familiar geometric figures. These are all examples of topological spaces known as manifolds. We investigate manifolds and some of the other ideas introduced in this section in more detail in Chapter 14.

EXAMPLE 3.21. Identify the opposite edges of a square, with the left and right edges glued as in Example 3.20, but now with the top edge glued to the bottom edge so that the left end of the top edge glues to the right end of the bottom edge, and the right end of the top edge glues to the left end of the bottom

edge. In other words, the top edge is flipped before it is glued to the bottom edge. (See Figure 3.27.) The resulting topological space is called the **Klein bottle**, named for the mathematician Felix Klein (1849–1925).
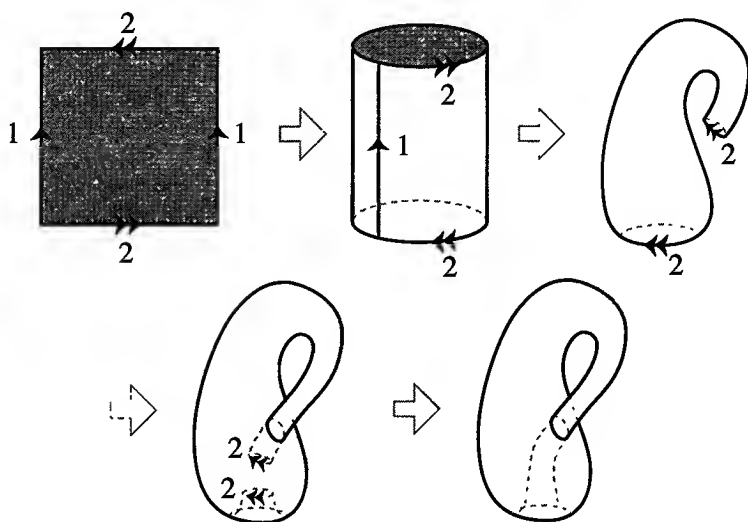
FIGURE 3.27: Gluing the edges of a square to obtain a representation of the Klein bottle.

In Figure 3.27, we show how to glue together a Klein bottle. We first glue the left and right edges of the square to create a cylinder. Now, when we attempt to glue the top circle of the cylinder to the bottom circle by bending the cylinder as we did for the torus, we find that the arrows on the two circles do not match. The only way to successfully match the arrows in 3-space is to allow the cylinder to pass through itself as if it were a soap film instead of a solid sheet of rubber. Then we can successfully glue the two circles together so that the arrows match. But of course, the actual Klein bottle does not have a circle of intersection with itself as this object does. In fact, the actual Klein bottle does not exist in 3-space. It can be constructed in 4-space without self-intersection. (See Chapter 11.) In 3-space, this representation with self-intersection is the best we can do.

EXAMPLE 3.22. Next we consider two more surfaces obtained as quotient spaces—one familiar, the other new and somewhat unusual. In both cases, we take the disk and divide its edge into two semicircles that we glue together. (See Figure 3.28 and 3.29.)

First, consider the quotient space obtained by the gluing shown in Figure 3.28. In that case, we are essentially zipping up a change purse to obtain a surface that is topologically equivalent to the sphere.
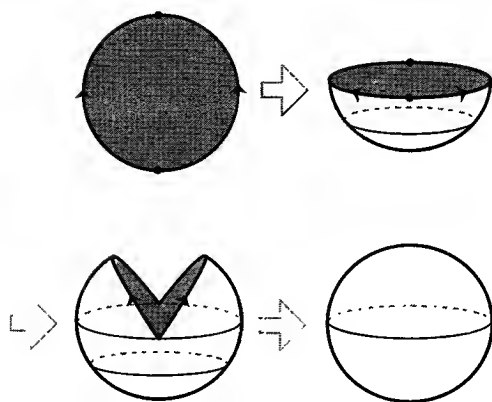
FIGURE 3.28: Gluing the boundary of a disk to obtain a sphere

Now consider the quotient space obtained by the gluing shown in Figure 3.29. The resulting topological space is known as the **projective plane**, denoted P. As in the Klein bottle example, we cannot construct the projective plane in 3-space. The best we can do to visualize the projective plane is to create a representation of it in 3-space, allowing self-intersection.
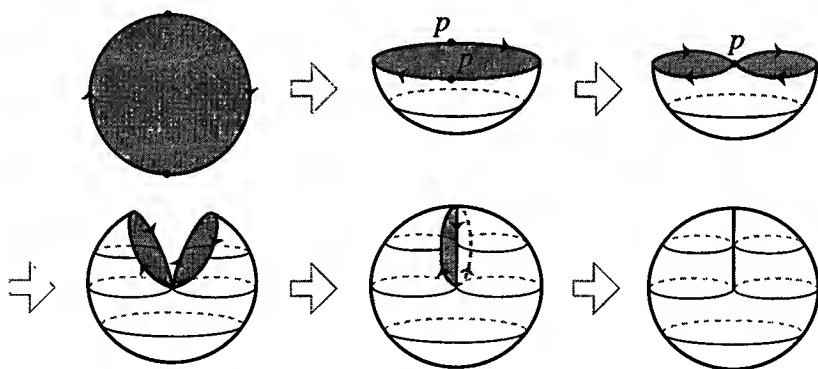
FIGURE 3.29: Gluing the boundary of a disk to obtain a representation of the projective plane.

In the illustrated construction of P, we begin by gluing together the points labeled p. The result is a bowl pinched along its rim. There are now two circles that need to be glued, but, unfortunately, as we bring them toward each other, we see that their arrows do not match properly, and we cannot directly glue them. We can, however, glue the back half of the left circle to the front half of the right circle, as we have done in the second-to-last stage in the illustration, leaving the other half of each circle unglued. To finish the gluing, we need to

pass part of the surface through itself. In the last stage in the illustration, we glue the remaining half circles together, and we depict it as if it is happening along the same vertical segment where the other two half circles are glued together. What is really happening, though, is that the surface is passing through itself along that segment so that the gluings are separate.

EXAMPLE 3.23. Consider the topological space obtained by identifying antipodal points (each pair of points x and $-x$) on the sphere. We show that the result is topologically equivalent to the projective plane.

With the equator as shown in Figure 3.30, note the following:

(i) Every point on the open northern hemisphere is identified with its antipode on the open southern hemisphere.

(ii) Every point on the equator is identified with its antipode on the equator.
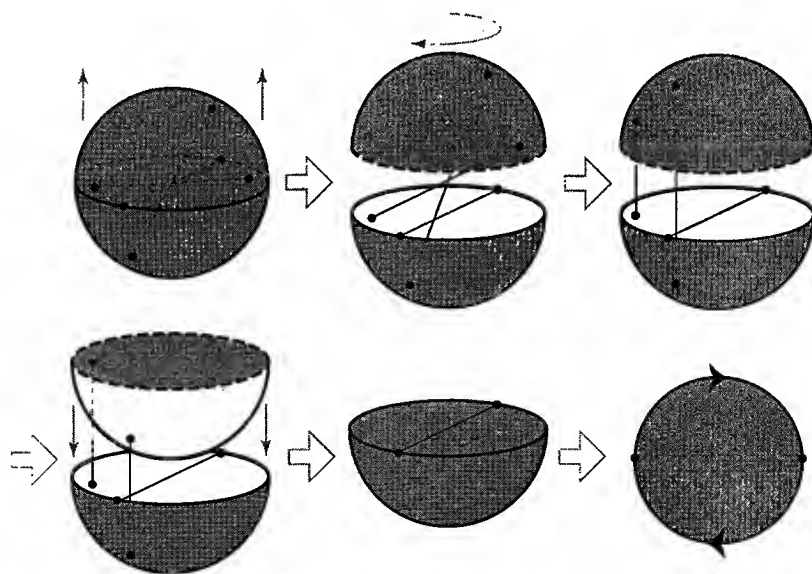
FIGURE 3.30: Identifying antipodal points on the sphere results in a projective plane.

We may remove the open northern hemisphere, rotate it, invert it, and then glue it directly onto the southern hemisphere as shown. Consequently, what remains is a bowl whose rim is glued together according to the arrows in Figure 3.30. That is equivalent to having a disk with the edge semicircles glued together as in Figure 3.29. Thus, the resulting space is topologically equivalent to the projective plane, as claimed.

Given two surfaces S_1 and S_2, there is a natural way to adjoin them to obtain a new surface. We describe it here, but leave the definition for Chapter 14. From both S_1 and S_2, remove the interior of a disk, leaving a hole surrounded by a circle. Then glue the two surrounding circles together via a quotient map. (See Figure 3.31.) The result is called the **connected sum** of S_1 and S_2 and is denoted $S_1 \# S_2$.
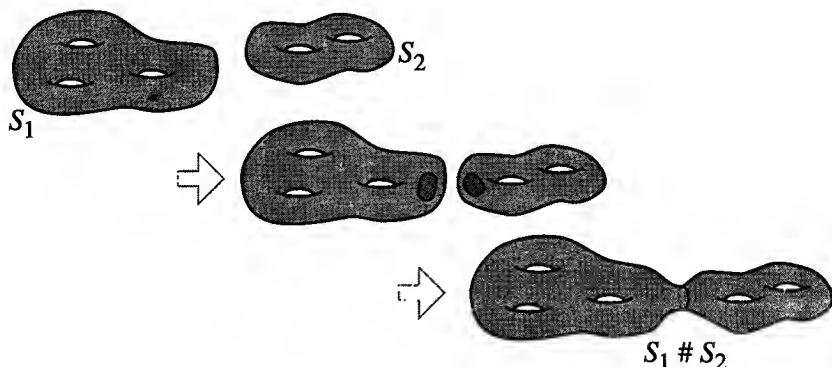
FIGURE 3.31: The connected sum of two surfaces.

EXAMPLE 3.24. In Figure 3.32 we show the surface, T # T, that results from taking the connected sum of two tori.
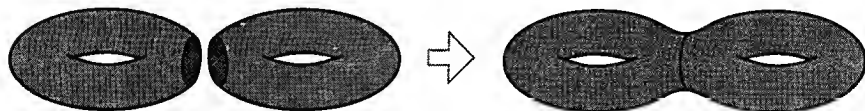
FIGURE 3 32: The connected sum of two tori.

By a process referred to as **cutting and pasting** we can obtain a representation of T # T as a polygon with edges glued together as illustrated in Figure 3.33. We start with two versions of the torus, represented as squares with edges glued together as in (a). Then we remove the interior of a disk from each as shown in (b). We choose those disks so that the boundary circle of each touches the vertex on the top edge and the right side edge of the corresponding square. This is for our convenience. Now, we open up those circles and stretch the squares out to pentagons, as shown in (c). For each of the resulting new edges, the two endpoints are still identified with each other, once we do all of the gluings on the edges. So in reality, each of these edges remains a circle in each surface. Next, we glue these two new edges together. This is the equivalent of taking the connected sum of two tori, since we are gluing together the boundaries of the two disks whose interiors we removed. Thus, in (d) we have $T \# T$, represented as an octagon with pairs of edges glued together.
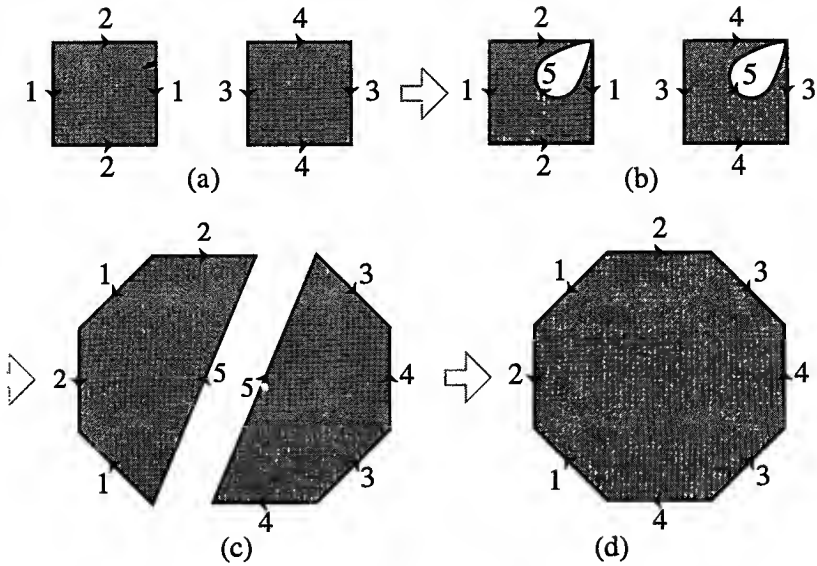
FIGURE 3.33: The connected sum of two tori is obtained by gluing edges in pairs on an octagon.

EXAMPLE 3.25. What do we obtain when we take the connected sum of two projective planes? In Figure 3.34, through cutting and pasting, we see that at step (e), $P\#P$ can be represented as a square with its edges glued together. If we
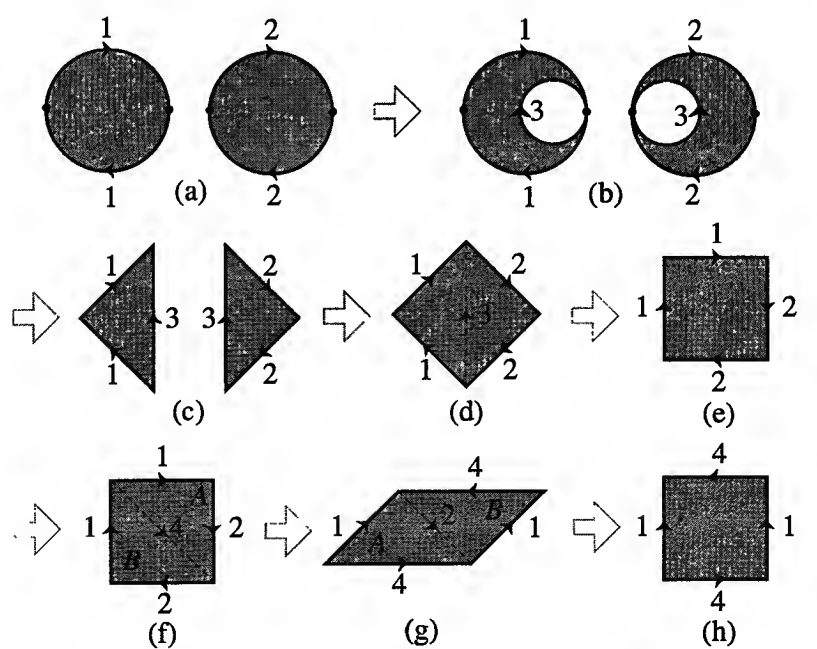
FIGURE 3.34: The connected sum of two projective planes is topologically equivalent to a Klein bottle.

cut across this square as in (f) and glue the edges labeled 2, then the result is a square with its edges glued together to yield a Klein bottle, as in Example 3.21. Thus, the connected sum of two projective planes is topologically equivalent to a Klein bottle.

In Figure 3.35 we show, through cutting and pasting, that removing the interior of a disk from a projective plane results in a Möbius band. Therefore, taking the connected sum of two projective planes is the same as gluing two Möbius bands together along their edges. The result is topologically equivalent to a Klein bottle.
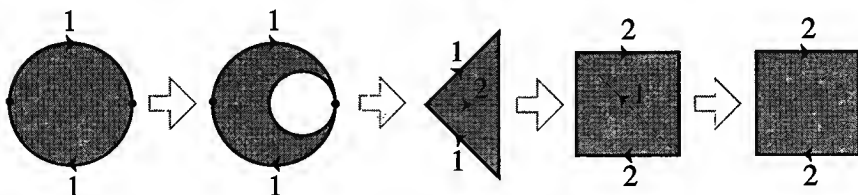
FIGURE 3.35: Removing the interior of a disk from a projective plane results in a Möbius band.

EXAMPLE 3.26. Consider the quotient space X obtained by identifying opposite faces of a cube $I \times I \times I$, as shown on the far left in Figure 3.36. We cannot build the corresponding object in 3-space, but the result of the gluing is a well-defined topological space. We can recognize the resulting topological space as follows: Glue the front and back identified faces together as in Figure 3.36. We obtain $S^1 \times I \times I$, a solid torus, with the top and bottom annuli and the inside and outside annuli identified as shown. In $S^1 \times I \times I$, at each point on the circle S^1, we have a square $I \times I$. In the quotient space X, the identifications on the top and bottom annuli and on the inside and outside annuli correspond to identifying the top and bottom edges and the side two edges, respectively, of each of these squares, as shown. The result of the identifications on each square is a torus. Therefore, we can view X as a product of a circle S^1 and a torus T. Since the torus is $S^1 \times S^1$, it follows that X is topologically equivalent to $S^1 \times S^1 \times S^1$, the 3-torus.
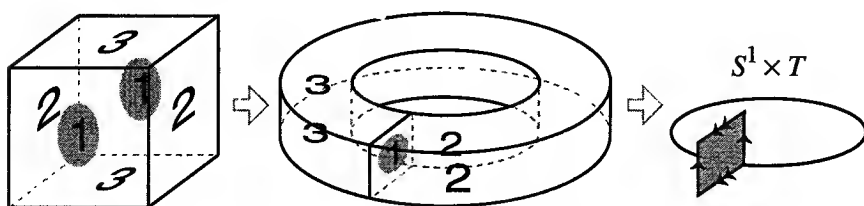
FIGURE 3.36: Identifying opposite faces of the cube as shown results in the 3-torus.

EXAMPLE 3.27. Consider the quotient space obtained by identifying opposite faces of a cube, as shown in Figure 3.37. If we take a line segment running from the center of the left face of the cube to the center of the right face, then when we glue the space together, the segment becomes a circle. If we take a cylinder centered on that segment, then when we glue the space together, the circles at the ends of the cylinder are glued together with a flip, resulting in a Klein bottle. Therefore, within the space there is a collection of concentric Klein bottles that shrinks down to a core circle.
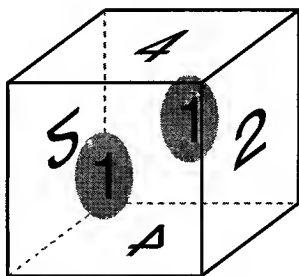
FIGURE 3.37: Gluing the faces of a cube to obtain $S^1 \times P$.

In Exercise 3.38 we ask you to show that this quotient space is topologically equivalent to $S^1 \times P$, the product of a circle and a projective plane.

The spaces in Examples 3.26 and 3.27 are examples of 3-manifolds. We investigate 3-manifolds further in Section 14.3, and, in Sections 14.4 and 14.5, we discuss which 3-manifolds could possibly correspond to the overall shape of the universe.

Exercises for Section 3.4

3.35. On a sketch of the surface T # T, illustrate where the glued edges of the octagon in Figure 3.33 appear.

3.36. (a) Show that a hexagon with opposite edges glued together straight across yields a torus.
(b) *Show that a hexagon with opposite edges glued together with a flip yields a projective plane.*

3.37. Give a representation of $T\#P$, the connected sum of a torus and a projective plane, as a hexagon with pairs of edges glued together.

3.38. Show that the quotient space in Example 3.27 is topologically equivalent to $S^1 \times P$, the product of a circle and a projective plane.

3.39. Show how to define a quotient space on a cube that results in $S^1 \times K$, the product of a circle and a Klein bottle. Verify that the gluing of your cube yields $S^1 \times K$.

3.5 *Configuration Spaces and Phase Spaces*

The connections between topology and physics run deep. It would take many volumes to cover all of the applications of topology to physics and it would require an advanced knowledge of topology to understand the most sophisticated applications. In many ways, there has been a strong interplay between the two fields—many topological concepts have been motivated by the study of physics applications, and many physics applications have been best understood through topological concepts.

The most basic topological construction in physics is that of the **configuration space**. In studying a system, we often need to keep track of a set of variables associated with the position and arrangement of a collection of objects of interest. For instance, we might want to keep track of the positions and orientations of the various parts of an industrial robot arm. The configuration space is a topological space that enables us to accomplish this variable tracking. Along with position, we may also want to consider the velocity or momentum of the objects in a system. In that case we use a space that includes both position and momentum variables; such a space is referred to as a **phase space** for the system.

In this section, we look at a few examples of configuration spaces and phase spaces, and we show how some of the topological spaces and constructions that we have already introduced arise naturally in applications.

EXAMPLE 3.28. To begin, consider the simple system where we have a rod with one end pinned at a point in the plane, around which the rod is free to rotate. (See Figure 3.38.) To each position of the rod, we associate a point on a circle centered at the pivot point, and to each such point on the circle, we associate a corresponding unique position of the rod. Thus the positions of the rod in the plane are modeled by the configuration space S^1.
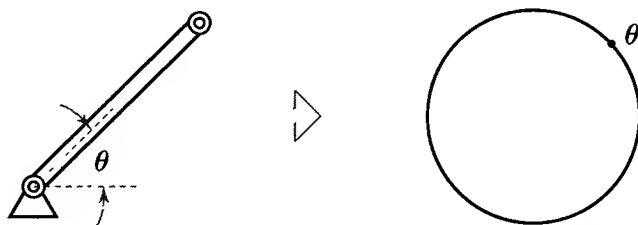
FIGURE 3.38: The configuration space of a rotating rod with one end pinned in the plane is a circle.

EXAMPLE 3.29. Now consider the planar two-rod system pictured in Figure 3.39. Rod A has one free end and one end that is pinned in the plane at a point around which the rod can rotate. Rod B is linked to the free end of rod A at a joint that allows rotation. Since the pinned end of rod A and the linked end of rod B each allow rotation through a full circle of positions, the configuration
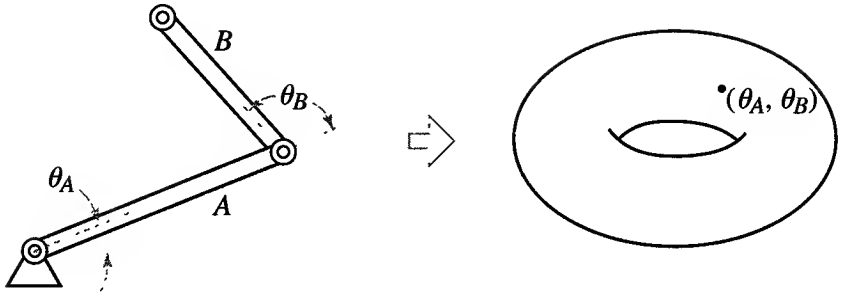
FIGURE 3.39: The configuration space of the two-rod system is the torus.

space of the system is $S^1 \times S^1$, which is the torus. For each point (θ_A, θ_B) on the torus, there is a configuration of our system, and each configuration of our system corresponds to a unique point on the torus.

It is important to avoid confusing the configuration space in Example 3.29 with the space in the plane that is swept out by the free end of the system. (See Figure 3.40.) Imagine putting a pen at the end of rod B and observing what space the pen traces as we run through all possible configurations of our system. If rod B and rod A are not equal in length, then the pen traces out an annulus in the plane. When the length of rod A equals the length of rod B, the pen traces out a disk. This "traced out" space is referred to as the **operational space** of the system. We explore the relationship between the configuration space and the operational space further in Section 4.3.
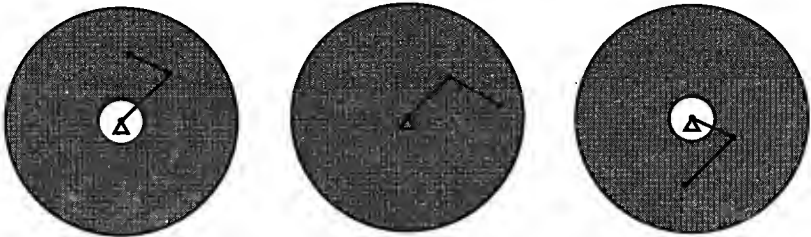
FIGURE 3.40: The operational space of the two-rod system is either an annulus or a disk.

The system in Example 3.29 is known as a **linkage**. A linkage is a collection of rigid parts that are connected together at joints where the parts are free to move relative to each other. In the eighteenth and nineteenth centuries, linkages were of mathematical interest because of their geometric properties. In the latter half of the nineteenth century, with the dawn of the industrial revolution, linkages became of applied interest as well, providing important components in machine design. Linkages are still employed in many mechanical systems. For example, a linkage converts the rotary motion of a motor into the back-and-forth motion of a windshield wiper. Linkages also continue to be of mathematical interest, generating problems explored in a variety of areas of mathematics. (For example, see [Con] and [Thu2].)

As manufacturing processes have become more complex and automated, so have the mechanical components involved. Robot arms are often employed to perform repetitive tasks that are done more cheaply, more accurately, and more precisely by a robot than by a human. In the next example and in Exercise 3.40, we examine configuration spaces associated with robot arms.

EXAMPLE 3.30. In Figure 3.41 we depict a robot arm that can vary along two angles and one length.
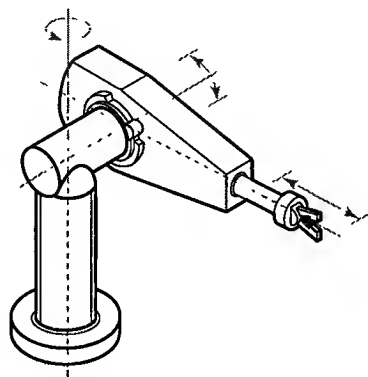
FIGURE 3.41: A robot arm

The base of the robot arm can turn a full 360 degrees. The corresponding configuration-space coordinate is a point on the circle S^1. The arm can turn from 0 to 90 degrees and therefore the corresponding configuration-space coordinate is constrained to lie in the interval $[0, 90]$. The arm can extend from 0 to 25 centimeters, so its configuration-space coordinate is constrained to lie in the interval $[0, 25]$. Thus, the configuration space of this robot arm is $S^1 \times [0, 90] \times [0, 25]$, as illustrated in Figure 3.42. This is topologically equivalent to $S^1 \times I \times I$, a solid torus.
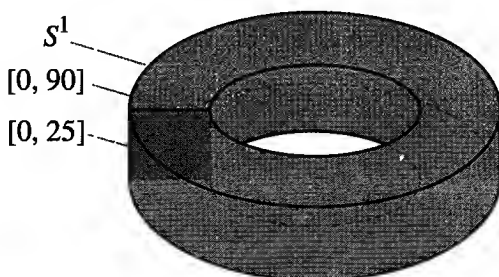
FIGURE 3.42: The configuration space for the robot arm is a solid torus.

EXAMPLE 3.31. Here we look at a simple application to sheet-metal folding. Suppose we have a square of sheet metal that we can fold along the horizontal
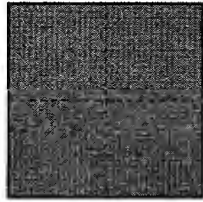
FIGURE 3.43: The square can be folded along the horizontal axis and the vertical axis.

axis and the vertical axis as shown in Figure 3.43. Let θ_H be the fold angle of the top half of the square with respect to the bottom half, where $\theta_H = 0$ represents a fold all of the way over to the front, $\theta_H = 2\pi$ represents a fold all of the way to the back, and $\theta_H = \pi$ corresponds to the unfolded situation as in the figure. Similarly, we let θ_V be the fold angle of the left half of the square with respect to the right half.

We make the physically reasonable assumption that we can fold along the vertical axis only when the square is folded along the horizontal axis with $\theta_H = 0, \pi$, or 2π. We make the corresponding assumption for folding along the horizontal axis. The configuration space of possible pairs of folding angles (θ_H, θ_V) is the subspace of the $\theta_H\theta_V$-plane, made up of three horizontal segments and three vertical segments, as illustrated in Figure 3.44.
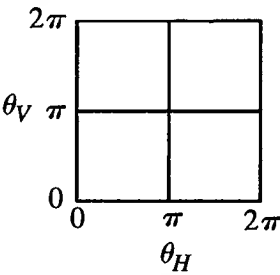
FIGURE 3.44: The configuration space of possible pairs of fold angles (θ_H, θ_V).

The bottom segment in the figure corresponds to varying the fold angle along the horizontal axis while the left half of the square is folded all of the way over in front of the right half. The other segments in the figure can be interpreted in like fashion.

EXAMPLE 3.32. In this example, we consider a bead in motion on an infinite
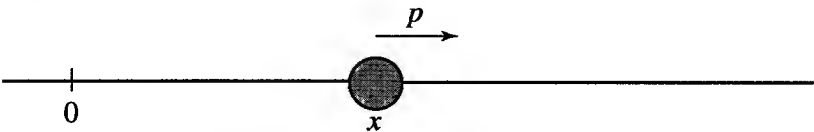
FIGURE 3.45: A bead in motion on an infinite wire.

wire, as illustrated in Figure 3.45. In this case, not only are we interested in the bead's position, but we also wish to track its momentum. Thus, two variables are needed, $x \in \mathbb{R}$ for position, and $p \in \mathbb{R}$ for momentum. Here we have a phase space rather than a configuration space, and the phase space is $\{(x, p) \mid x, p \in \mathbb{R}\}$. This, of course, is just the product space $\mathbb{R} \times \mathbb{R} = \mathbb{R}^2$.

EXAMPLE 3.33. Suppose an object of mass m is falling or rising vertically, near the surface of the earth, under the influence of gravity. Let x be the height above the ground and p be the momentum of the object. Then the phase space is given by $\{(x, p) \mid x, p \in \mathbb{R}, \; x \geq 0\} = \mathbb{R}_+ \times \mathbb{R}$. For motion within this system, assume that the total energy is conserved. The total energy is expressed as the sum of the potential energy, mgx (where g represents the acceleration due to gravity) and the kinetic energy, $\frac{p^2}{2m}$. Then, when we examine phase space tracks of the changing position and momentum of our object, we find that the tracks must remain restricted to a subspace of $\mathbb{R}_+ \times \mathbb{R}$ of the form

$$E_C = \{(x, p) \in \mathbb{R}_+ \times \mathbb{R} \mid mgx + \tfrac{p^2}{2m} = C\},$$

for $C \geq 0$. (See Figure 3.46.) The sets E_C partition the phase space into parabolic curves of constant energy along which all phase space tracks of the system's motion must lie.
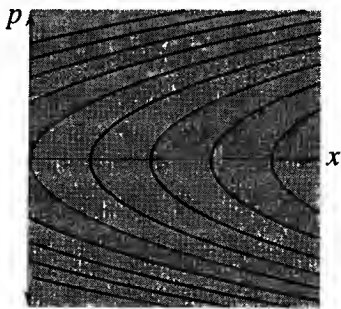
FIGURE 3.46: Constant energy curves in the phase space of a falling or rising object.

Often, in the study of a given system, conserved quantities (such as total energy in the previous example) define subspaces of the configuration space or phase space that aid in the investigation of the system.

EXAMPLE 3.34. Suppose we have a molecule, M, that consists of two distinct atoms, and we wish to record its position and orientation in 3-space. (See Figure 3.47.) We associate to the molecule the point in $\mathbb{R}^3$ at its center of mass. This enables us to locate the molecule in 3-space, but we need more information to determine the orientation of the molecule once its center of mass has
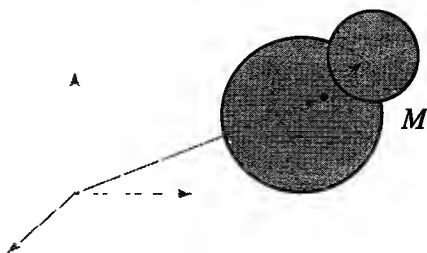
FIGURE 3.47: An oriented molecule in 3-space.

been situated. This can be accomplished by associating a point on the 2-sphere to each possible orientation, as follows:

We pick one of the two atoms, and to each orientation of the molecule, we associate a vector pointing from the center of mass to the center of the chosen atom. To that vector (and therefore that orientation) we associate the point on the sphere to which the vector points when placed with its tail at the center of the sphere. In this way, each orientation corresponds to a different point on the sphere. Hence, the configuration space for recording the position and orientation of the molecule in 3-space is $\mathbb{R}^3 \times S^2$.

EXAMPLE 3.35. Now suppose that the molecule in Example 3.34 is made of two identical atoms. Again, every point in $\mathbb{R}^3$ corresponds to a possible location of the molecule's center of mass. Also, at each location in $\mathbb{R}^3$, points on the sphere correspond to each possible orientation of the molecule. However, because of the symmetry of the molecule, each pair of antipodal points in the sphere corresponds to indistinguishable configurations. Thus we can "reduce" the configuration space by identifying antipodal points on each sphere of molecule orientations. Therefore the configuration space is $\mathbb{R}^3 \times P$, where P is the projective plane.

As in Example 3.35, it is common in physics and chemistry to obtain a configuration space or a phase space as a quotient space of another space, taking advantage of symmetry properties of the system.

Exercises for Section 3.5

3.40. Determine the configuration space for each of the following three robot arms, which are illustrated in Figure 3.48:

(a) The robot arm has two components that can rotate through a full circle. The whole robot arm can be translated along a base track that is 50 centimeters long. The height of the robot arm can be raised 20 centimeters from its lowest position.

(b) The robot arm has four moving components that each can rotate through a full circle.

(c) The robot arm has two moving components that each can rotate through a full circle. The arm of the robot can rotate 90 degrees from a horizontal position to a vertical position.
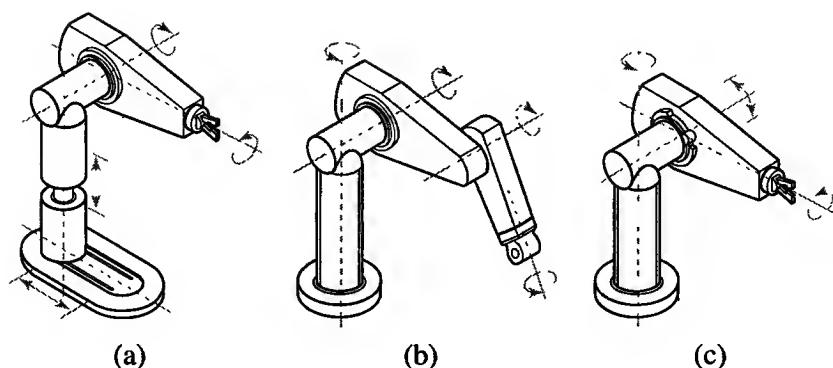
FIGURE 3.48: A variety of robot arms.

3.41. Consider the system shown in Figure 3.49. It consists of a rod with one end free and one end attached to a track that is one meter long. The point of attachment can slide along the track, and the rod can rotate through a full circle at the point of attachment. Determine the configuration space of the system.
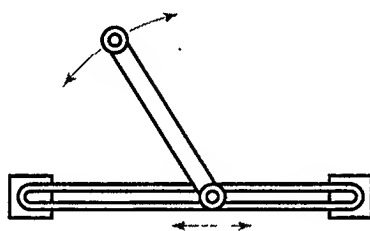
FIGURE 3.49: A rotating rod on a track.

3.42. Suppose we have a disk-shaped piece of sheet metal that can be folded along three axes as shown in Figure 3.50. Describe and illustrate the configuration space of possible fold angles for the disk.
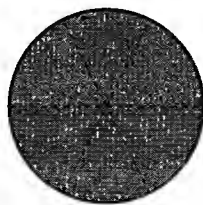
FIGURE 3.50: The disk can be folded along each of the axes shown.

3.43. Describe the phase space (position and momentum) for the motion of a bead on a circular wire.

3.44. Determine the configuration space for the position and orientation of a water molecule, H_2O, in the plane.

Chapter 4

Continuous Functions and Homeomorphisms

Now that we have defined various concepts related to topological spaces and their subsets, we would like to consider functions sending one topological space to another. The concept of continuity of functions is one of the most important in topology.

A topology on a set is a structure that establishes a notion of proximity on the set. Continuous functions between topological spaces preserve proximity, reflecting the idea that a continuous function sends points that are close in one space to points that are close in the other. In Section 4.1, we give a topological definition of continuity and we present results and examples related to continuity.

A continuous bijective function that has a continuous inverse is called a homeomorphism. Such functions provide us with the main notion of topological equivalence. In Section 4.2, we introduce and explore homeomorphisms and topological equivalence.

We close this chapter in Section 4.3 with a discussion of the forward kinematics map, a continuous function that plays an essential role in the study of mechanical systems.

4.1 Continuity

Our first formal exposure to continuity usually comes in a calculus or analysis course, focusing on functions mapping the real line $\mathbb{R}$ to itself. Typically, it is defined as follows:

DEFINITION 4.1. *A function $f : \mathbb{R} \to \mathbb{R}$ is **continuous** if for every $x_0 \in \mathbb{R}$ and every $\varepsilon > 0$, there exists a $\delta > 0$ such that if $|x - x_0| < \delta$, then $|f(x) - f(x_0)| < \varepsilon$.*

In other words, f is continuous if for every x_o and ε, we can find a δ so that if x is close enough to x_0 (within δ), then $f(x)$ is close to $f(x_0)$ (within ε). This is often called the $\varepsilon - \delta$ **definition of continuity**. It is an extremely useful concept that plays a pivotal role in analysis.

We provide a general definition of continuity for functions that map from one topological space to another. This topological definition of continuity is very simple to state and, as we will show, is equivalent to the $\varepsilon - \delta$ definition for functions that map $\mathbb{R}$ to $\mathbb{R}$.

DEFINITION 4.2. *Let X and Y be topological spaces. A function $f : X \to Y$ is **continuous** if $f^{-1}(V)$ is open in X for every open set V in Y.*

We call this the **open set definition of continuity**. Paraphrased, it states that f is continuous if the preimage of every open set is open.

EXAMPLE 4.1. Let $X = \{a, b, c, d\}$ and $Y = \{1, 2, 3\}$ have the topologies as depicted in Figure 4.1. Let functions $f, g, h : X \to Y$ be defined by

$$f(a) = 1, \ f(b) = 1, \ f(c) = 2, \ f(d) = 2,$$

$$g(a) = 2, \ g(b) = 2, \ g(c) = 1, \ g(d) = 3,$$

$$h(a) = 1, \ h(b) = 2, \ h(c) = 2, \ h(d) = 3.$$

The function f is continuous, as can be easily verified by checking that the preimage of each open set in Y is open in X. Similarly, g is continuous. However, h is not continuous because $\{2\}$ is open in Y, but $h^{-1}(\{2\}) = \{b, c\}$ is not open in X.
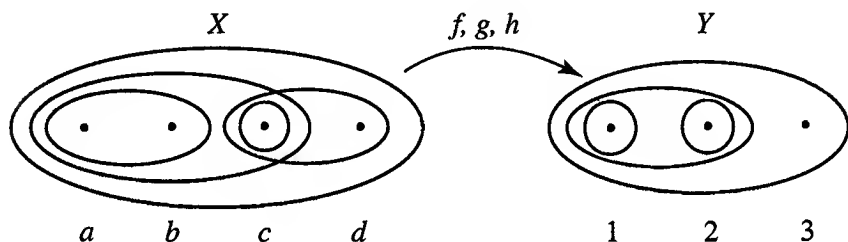
FIGURE 4.1: Functions f, g, and h map X to Y.

EXAMPLE 4.2. Let X and Y be topological spaces.

   (i) The identity function $id : X \to X$, given by $id(x) = x$, is continuous.

   (ii) Pick $y_0 \in Y$, and consider the constant function, $C : X \to Y$, defined by $C(x) = y_0$ for every $x \in X$. We show that C is continuous. Suppose that V is open in Y; then $C^{-1}(V) = X$ if $y_0 \in V$, and $C^{-1}(V) = \varnothing$ if $y_0 \notin V$. In either case $C^{-1}(V)$ is open in X, and therefore C is continuous.

It is not just the formula or the relation defining a function that determines whether or not the function is continuous; the topology on the domain and the topology on the range are also significant. Consider the following example.

EXAMPLE 4.3. Let $f : \mathbb{R} \to \mathbb{R}_{fc}$ be given by $f(x) = \frac{x-1}{2}$, where the domain $\mathbb{R}$ has the standard topology and the range $\mathbb{R}_{fc}$ is the real line in the finite complement topology. The function f is continuous since if

$V = \mathbb{R} - \{x_1, \ldots, x_m\}$ is an open set in the finite complement topology, then $f^{-1}(V) = \mathbb{R} - \{2x_1 + 1, \ldots, 2x_m + 1\}$, and this is an open set in $\mathbb{R}$ with the standard topology.

Now let $g : \mathbb{R} \to \mathbb{R}_l$ be given by $g(x) = \frac{x-1}{2}$, where the domain $\mathbb{R}$ has the standard topology and the range $\mathbb{R}_l$ is the real line in the lower limit topology. The function g is expressed by the same formula as f, but g is not continuous. The set $[0, 1)$ is open in $\mathbb{R}_l$, but $g^{-1}([0, 1)) = [1, 3)$ is not open in $\mathbb{R}$ with the standard topology.

EXAMPLE 4.4. Let X be a topological space and Y be a subspace of X. The inclusion function $i : Y \to X$, given by $i(y) = y$, is continuous. Why? If we take an open set U in X, then $i^{-1}(U) = U \cap Y$, and $U \cap Y$ is an open set in Y by the definition of the subspace topology on Y. In fact, the subspace topology on Y is the coarsest topology on Y for which the inclusion function i is continuous. (See Exercise 4.5(a).)

Let $p : X \to A$ be a surjective function from a topological space X to a set A, and let A have the quotient topology induced by p. Then p is continuous. Why? By the definition of the quotient topology, a set V is open in A if and only if $p^{-1}(V)$ is open in X. Therefore, the preimage of every open set in A is automatically open in X. In fact, the quotient topology on A is the finest topology on A for which p is continuous. (See Exercise 4.5(b).)

By definition, a function $f : X \to Y$ is continuous if the preimage of every open set in Y is open in X. However, checking that every open set in Y has an open preimage in X is more than we really need to do. As the following theorem indicates, to prove that f is continuous, it suffices to consider only the sets in a fixed basis for Y, showing that the preimage of each basis element is open in X.

THEOREM 4.3. *Let X and Y be topological spaces and $\mathcal{B}$ be a basis for the topology on Y. Then $f : X \to Y$ is continuous if and only if $f^{-1}(B)$ is open in X for every $B \in \mathcal{B}$.*

    Proof. Suppose $f : X \to Y$ is continuous. Then $f^{-1}(V)$ is open in X for every V open in Y. Since every basis element B is open in Y, it follows that $f^{-1}(B)$ is open in X for all $B \in \mathcal{B}$.

    Now, suppose $f^{-1}(B)$ is open in X for every $B \in \mathcal{B}$. We show that f is continuous. Let V be an open set in Y. Then V is a union of basis elements, say $V = \cup B_\alpha$. Thus,

$$f^{-1}(V) = f^{-1}(\cup B_\alpha) = \cup f^{-1}(B_\alpha).$$

    By assumption, each set $f^{-1}(B_\alpha)$ is open in X; therefore so is their union. Thus, $f^{-1}(V)$ is open in X, and it follows that the preimage of every open set in Y is open in X. Hence, f is continuous. ∎

EXAMPLE 4.5. The functions $f(x) = x + 2$, $g(x) = 2x$, and $h(x) = x^2$ are all continuous functions mapping $\mathbb{R}$ to $\mathbb{R}$ in the standard topology.

Why? Let (a, b), with $a < b$, be a basis element for the standard topology. Then

$$f^{-1}((a,b)) = (a - 2, b - 2),$$

$$g^{-1}((a,b)) = \left(\frac{a}{2}, \frac{b}{2}\right),$$

$$h^{-1}((a,b)) = \begin{cases} (-\sqrt{b}, -\sqrt{a}) \cup (\sqrt{a}, \sqrt{b}) & \text{if } a \geq 0, \\ (-\sqrt{b}, \sqrt{b}) & \text{if } a < 0 \text{ and } b > 0, \\ \varnothing & \text{if } b \leq 0. \end{cases}$$

In each case, the preimage of an arbitrary basis element is an open set. Thus, each function is continuous.

EXAMPLE 4.6. Multiplication is a continuous function. That is, the function $f : \mathbb{R}^2 \to \mathbb{R}$, given by $f(x, y) = xy$, is continuous. To see this, take a basis element (a, b) in the range $\mathbb{R}$. Its preimage in $\mathbb{R}^2$ is the set

$$f^{-1}((a, b)) = \{(x, y) \in \mathbb{R}^2 \mid a < xy < b\}.$$

This is a set sandwiched between two hyperbolas. An example, with a and b both positive, is illustrated in Figure 4.2.
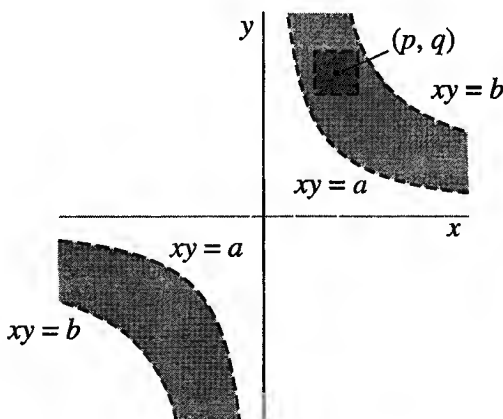
FIGURE 4.2: The preimage $f^{-1}((a, b)) = \{(x, y) \in \mathbb{R}^2 \mid a < xy < b\}$ with a and b both positive.

It is intuitively clear that $f^{-1}((a, b))$ is an open subset of $\mathbb{R}^2$. We can prove this explicitly by showing that for each point $(p, q) \in f^{-1}((a, b))$, there is an open square centered at (p, q) and contained in $f^{-1}((a, b))$. (See Figure 4.2.) In fact, if we set $m = \min\{b - pq, pq - a\}$ and choose $\delta > 0$ so that $\delta|p|$, $\delta|q|$, and $3\delta^2$ are all less than $m/3$, then it can be shown that $(p - \delta, p + \delta) \times (q - \delta, q + \delta) \subset f^{-1}((a, b))$. (See Exercise 4.15.)

The addition function on the real numbers is also continuous. (See Exercise 4.14.) Combining the continuity of multiplication and the continuity of addition, we can obtain the following theorem:

THEOREM 4.4. *Let $\mathbb{R}$ have the standard topology. Then every polynomial function $p : \mathbb{R} \to \mathbb{R}$, with $p(x) = a_n x^n + ... + a_1 x + a_0$, is continuous.*

Proof. See Exercises 4.13–4.17. ∎

The open set definition of continuity, and results based on it, can be used to prove that rational functions, trigonometric functions and their inverses, exponential functions, and logarithmic functions are all continuous on domains where they are defined. We do not prove these results here, but throughout the rest of the text we assume them.

The topological definition of continuity is very simple to state, but it might leave you wondering how the definition reflects the idea that a continuous function preserves proximity. The following theorem provides some insight, indicating that a continuous function maps a point in the closure of a set to a point in the closure of the image of the set.

THEOREM 4.5. *Let $f : X \to Y$ be continuous and assume that $A \subset X$. If $x \in Cl(A)$, then $f(x) \in Cl(f(A))$. (See Figure 4.3.)*
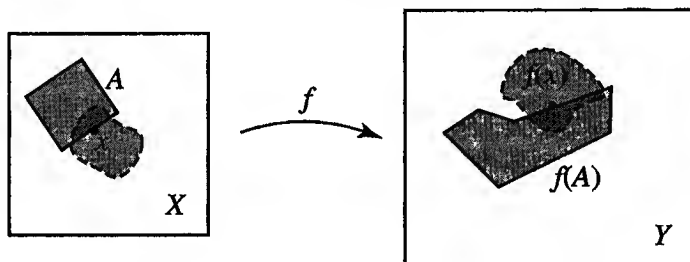
FIGURE 4.3: If $x \in Cl(A)$, then $f(x) \in Cl(f(A))$

Proof. Suppose that $f : X \to Y$ is continuous, $x \in X$, and $A \subset X$. We prove that if $f(x) \notin Cl(f(A))$, then $x \notin Cl(A)$. Thus suppose that $f(x) \notin Cl(f(A))$. By Theorem 2.5 there exists an open set U containing $f(x)$, but not intersecting $f(A)$. It follows that $f^{-1}(U)$ is an open set containing x that does not intersect A. Thus $x \notin Cl(A)$, and the result follows. ∎

Next we consider the equivalence between the open set definition of continuity and the $\varepsilon - \delta$ definition of continuity in the case of functions that map $\mathbb{R}$ to $\mathbb{R}$. A translation of the $\varepsilon - \delta$ definition into topological terms might read as follows:

Translation of the $\varepsilon - \delta$ Definition: Let X and Y be topological spaces. A function $f : X \to Y$ is continuous if, for every $x \in X$ and every open set U containing $f(x)$, there exists a neighborhood V of x such that $f(V) \subset U$. (See Figure 4.4.)
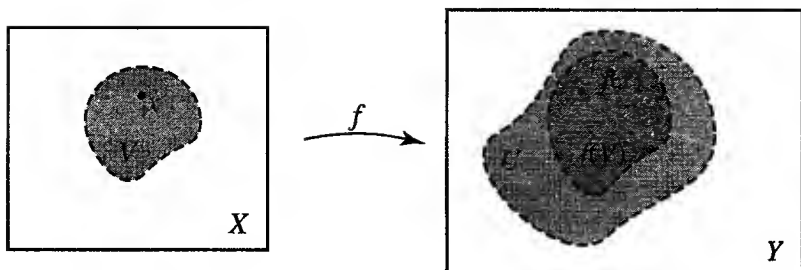
FIGURE 4.4: For every U containing $f(x)$, there exists V containing x such that $f(V) \subset U$.

In the translation, the open set U is playing the role of the ε-interval in the $\varepsilon - \delta$ definition, and the open set V is playing the role of the δ-interval. It is straightforward to show that for functions $f : \mathbb{R} \to \mathbb{R}$, the $\varepsilon - \delta$ definition of continuity is equivalent to the translation. (See Exercise 4.3.) In the next theorem, we show that the translation is equivalent to the open set definition.

THEOREM 4.6. *A function $f : X \to Y$ is continuous in the open set definition of continuity if and only if, for every $x \in X$ and every open set U containing $f(x)$, there exists a neighborhood V of x such that $f(V) \subset U$.*

Proof. First, suppose that the open set definition holds for functions $f : X \to Y$. Let $x \in X$ and an open set $U \subset Y$ containing $f(x)$ be given. Set $V = f^{-1}(U)$. It follows that $x \in V$ and that V is open in X since f is continuous by the open set definition. Clearly $f(V) \subset U$, and therefore we have shown the desired result.

Now assume that for every $x \in X$ and every open set U containing $f(x)$, there exists a neighborhood V of x such that $f(V) \subset U$. We show that $f^{-1}(W)$ is open in X for every open set W in Y. Thus let W be an arbitrary open set in Y. To show that $f^{-1}(W)$ is open in X, choose an arbitrary $x \in f^{-1}(W)$. It follows that $f(x) \in W$, and therefore there exists a neighborhood V_x of x in X such that $f(V_x) \subset W$, or, equivalently, such that $V_x \subset f^{-1}(W)$. Thus, for an arbitrary $x \in f^{-1}(W)$ there exists an open set V_x such that $x \in V_x \subset f^{-1}(W)$. Theorem 1.4 implies that $f^{-1}(W)$ is open in X. ∎

Another sense in which continuity preserves proximity is demonstrated by the fact that continuous functions map convergent sequences to convergent sequences, as the following theorem indicates:

THEOREM 4.7. *Assume that $f : X \to Y$ is continuous. If a sequence $(x_1, x_2, \dots)$ in X converges to a point x, then the sequence $(f(x_1), f(x_2), \dots)$ in Y converges to $f(x)$.*

Proof. Let U be an arbitrary neighborhood of $f(x)$ in Y. Since f is continuous, $f^{-1}(U)$ is open in X. Furthermore, $f(x) \in U$ implies that $x \in f^{-1}(U)$. The sequence $(x_1, x_2, \dots)$ converges to x; thus, there

exists $N \in \mathbb{Z}_+$ such that $x_n \in f^{-1}(U)$ for all $n \geq N$. It follows that $f(x_n) \in U$ for all $n \geq N$, and therefore the sequence $(f(x_1), f(x_2), \ldots)$ converges to $f(x)$.   ■

IMPORTANT NOTE: *A continuous function does not necessarily map open sets to open sets. For example, the function $f : \mathbb{R} \to \mathbb{R}$, given by $f(x) = x^2$, is continuous, but the image of the open set $(-1, 1)$ is $[0, 1)$, which is not open.*

The following theorem indicates that the open set definition of continuity has an equivalent closed set version.

THEOREM 4.8. *Let X and Y be topological spaces. A function $f : X \to Y$ is continuous if and only if $f^{-1}(C)$ is closed in X for every closed set $C \subset Y$.*

   Proof. See Exercise 4.2.   ■

It is straightforward to show that compositions of continuous functions are continuous. Specifically, we have the following theorem:

THEOREM 4.9. *Let $f : X \to Y$ and $g : Y \to Z$ be continuous. Then the composition function, $g \circ f : X \to Z$, is continuous.*

   Proof. Suppose that $f : X \to Y$ and $g : Y \to Z$ are continuous, and let U be an open set in Z. Then $(g \circ f)^{-1}(U) = f^{-1}(g^{-1}(U))$. Since g is continuous, $g^{-1}(U)$ is open in Y, and since f is continuous, $f^{-1}(g^{-1}(U))$ is open in X. Thus, $(g \circ f)^{-1}(U)$ is open in X for an arbitrary U open in Z, implying that $g \circ f$ is continuous.   ■

We will find the following lemma useful. It is called the Pasting Lemma because it provides conditions under which we can paste together continuous functions $f : A \to Y$ and $g : B \to Y$ to obtain a continuous function, $h : A \cup B \to Y$, defined on the union of the domains of f and g. (See Figure 4.5.)
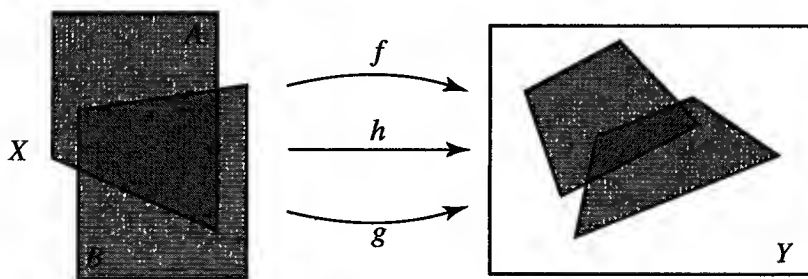
FIGURE 4.5: The functions $f : A \to Y$ and $g : B \to Y$ paste together to define $h : A \cup B \to Y$.

LEMMA 4.10. *The Pasting Lemma.* *Let X be a topological space and let A and B be closed subsets of X such that $A \cup B = X$. Assume that $f : A \to Y$ and $g : B \to Y$ are continuous and $f(x) = g(x)$ for all x in $A \cap B$. Then $h : X \to Y$, defined by*

$$h(x) = \begin{cases} f(x) & \text{if } x \in A, \\ g(x) & \text{if } x \in B, \end{cases}$$

is a continuous function.

Proof. By Theorem 4.8, it suffices to show that if C is closed in Y, then $h^{-1}(C)$ is closed in X. Thus suppose that C is closed in Y. Note that $h^{-1}(C) = f^{-1}(C) \cup g^{-1}(C)$. Since f is continuous, it follows by Theorem 4.8 that $f^{-1}(C)$ is closed in A. Theorem 3.4 then implies that $f^{-1}(C) = D \cap A$ where D is closed in X. Now, D and A are both closed in X, and $f^{-1}(C) = D \cap A$; therefore, $f^{-1}(C)$ is closed in X. Similarly, $g^{-1}(C)$ is closed in X. Thus, $h^{-1}(C)$ is the union of two closed sets in X and therefore is closed in X as well. It follows that h is continuous. ∎

EXAMPLE 4.7. The absolute value function $h : \mathbb{R} \to \mathbb{R}$, given by

$$h(x) = |x| = \begin{cases} x & \text{if } x \geq 0, \\ -x & \text{if } x \leq 0, \end{cases}$$

consists of two continuous functions defined on closed subsets of $\mathbb{R}$. One is the function $f : (-\infty, 0] \to \mathbb{R}$, given by $f(x) = -x$, and the other is the function $g : [0, \infty) \to \mathbb{R}$, given by $g(x) = x$. These two functions are pieced together at $x = 0$, where they agree. Hence, the resulting function h is continuous.

Exercises for Section 4.1

4.1. (a) Let X have the discrete topology and Y be an arbitrary topological space. Show that every function $f : X \to Y$ is continuous.

(b) Let Y have the trivial topology and X be an arbitrary topological space. Show that every function $f : X \to Y$ is continuous.

4.2. **Prove Theorem 4.8:** Let X and Y be topological spaces. A function $f : X \to Y$ is continuous if and only if $f^{-1}(C)$ is closed in X for every closed set $C \subset Y$.

4.3. Show that a function $f : \mathbb{R} \to \mathbb{R}$ is continuous in the $\varepsilon - \delta$ definition of continuity if and only if, for every $x \in \mathbb{R}$ and every open set U containing $f(x)$, there exists a neighborhood V of x such that $f(V) \subset U$.

4.4. Prove whether or not each given function is continuous.

(a) The function $f : \mathbb{R}_l \to \mathbb{R}$, defined by $f(x) = 3x - 5$, where $\mathbb{R}_l$ is the real line with the lower limit topology.

(b) The function $g : \mathbb{R}_{fc} \to \mathbb{R}$, defined by $g(x) = 3x - 5$, where $\mathbb{R}_{fc}$ is the real line with the finite complement topology.

4.5. (a) Let X be a topological space, and let Y be a subset of X. Show that the subspace topology on Y is the coarsest topology on Y for which the inclusion function, $i : Y \to X$, defined by $i(y) = y$, is continuous. That is, show that if $\mathcal{T}$ is the subspace topology on Y, and $\mathcal{T}'$ is a topology on Y such that i is continuous, then $\mathcal{T} \subset \mathcal{T}'$.

(b) Let $p : X \to A$ be a surjective function from a topological space X to a set A. Show that the quotient topology on A induced by p is the finest topology on A for which the function p is continuous. That is, show that if $\mathcal{T}$ is the quotient topology on A, and $\mathcal{T}'$ is a topology on A such that p is continuous, then $\mathcal{T}' \subset \mathcal{T}$.

4.6. (a) Let X and Y be topological spaces, and let $X \times Y$ be the corresponding product space. Define the projection functions

$$p_X : X \times Y \to X \text{ and } p_Y : X \times Y \to Y$$

by $p_X(x, y) = x$ and $p_Y(x, y) = y$. Prove that p_X and p_Y are continuous.

(b) Show that the product topology on $X \times Y$ is the coarsest topology on $X \times Y$ for which both functions p_X and p_Y are continuous. That is, show that if $\mathcal{T}$ is the product topology on $X \times Y$, and $\mathcal{T}'$ is a topology on $X \times Y$ such that p_X and p_Y are continuous, then $\mathcal{T} \subset \mathcal{T}'$.

4.7. Suppose X is a space with topologies $\mathcal{T}_1$ and $\mathcal{T}_2$. Let $id : X \to X$ be the identity, $id(x) = x$, and assume that the domain X has topology $\mathcal{T}_1$ and that the range X has topology $\mathcal{T}_2$. Show that id is continuous if and only if $\mathcal{T}_1$ is finer than $\mathcal{T}_2$.

4.8. Let $f : X \to Y$ be a continuous function. If x is a limit point of a subset A of X, is it true that $f(x)$ is a limit point of $f(A)$ in Y? Prove this or find a counterexample.

4.9. Let $f, g : X \to Y$ be continuous functions. Assume that Y is Hausdorff and that there exists a dense subset D of X such that $f(x) = g(x)$ for all $x \in D$. Prove that $f(x) = g(x)$ for all $x \in X$.

4.10. Let $f : X \to Y$ be a function. The graph of f is the subset of $X \times Y$ given by $G = \{(x, f(x)) \mid x \in X\}$. Show that, if f is continuous and Y is Hausdorff, then G is closed in $X \times Y$. (Note: In Exercise 7.13 we consider a converse of this result, assuming Y also satisfies a property known as compactness.)

4.11. Let $f : X \to Y$ be continuous, and let A be a subspace of X. Prove that $f|_A : A \to Y$, the restriction of f to A, is continuous.

4.12. (a) Prove that the Pasting Lemma holds if we replace the assumption that A and B are both closed in X with the assumption that A and B are both open in X.

(b) Provide an example showing that the Pasting Lemma does not hold if we drop the assumption that A and B are both closed in X.

4.13. (a) Let $f_1 : X \to Y_1$ and $f_2 : X \to Y_2$ be continuous functions. Show that $h : X \to Y_1 \times Y_2$, defined by $h(x) = (f_1(x), f_2(x))$, is continuous as well.

(b) Extend the result of (a) to n functions, for $n > 2$.

4.14. Show that the addition function, $f : \mathbb{R}^2 \to \mathbb{R}$, given by $f(x, y) = x + y$, is a continuous function.

4.15. Let f be the multiplication function, $f(x, y) = xy$. Complete the proof of the continuity of f that was outlined in Example 4.6, by doing the following:

    (a) Show that if p and q are both positive, and δ is as described in the example, then $(p - \delta, p + \delta) \times (q - \delta, q + \delta) \subset f^{-1}((a, b))$.

    (b) Consider the rest of the possibilities for p and q being positive or negative, and show that $(p - \delta, p + \delta) \times (q - \delta, q + \delta) \subset f^{-1}((a, b))$. (Note: You can argue by symmetry here. For example, the situation for p and q both negative is symmetric about the origin to the situation for p and q both positive.)

4.16. Use Example 4.6, Exercises 4.13 and 4.14, and Theorem 4.9 to show that the sum and product of a finite number of continuous functions are also continuous functions. That is, assuming that $f_1, \ldots, f_m : \mathbb{R} \to \mathbb{R}$ are continuous, prove that $S : \mathbb{R} \to \mathbb{R}$ and $P : \mathbb{R} \to \mathbb{R}$, defined by $S(x) = f_1(x) + \ldots + f_m(x)$ and $P(x) = f_1(x)f_2(x)\ldots f_m(x)$, are continuous.

4.17. Use Exercise 4.16 to show that every polynomial function $p : \mathbb{R} \to \mathbb{R}$, given by $p(x) = a_n x^n + \ldots + a_1 x + a_0$, is continuous.

Supplementary Exercises: Sequences of Functions

Sequences of functions play an important role in topology and analysis. For example, functions that are defined by power series, such as $\sin(x)$ and e^x, can be shown to be continuous because they are limits of converging sequences of polynomials, which themselves are continuous functions. In this text, we employ a convergent sequence of functions in the proof of the Tietze Extension Theorem, which appears in a set of supplementary exercises in Chapter 7.

In these exercises, we consider conditions that ensure that a sequence of continuous functions converges to a continuous function. To begin, consider the sequence of functions (s_n) where, for each $n = 1, 2, 3, \ldots$, the function $s_n : [0, 1] \to [0, 1]$ is defined by

$$s_n(x) = \begin{cases} nx & \text{if } 0 \le x \le \frac{1}{n}, \\ 1 & \text{if } \frac{1}{n} \le x \le 1. \end{cases}$$

SE 4.18. Sketch graphs of s_1, s_2, and s_3, and prove that s_n is continuous for all $n = 1, 2, 3, \ldots$.

DEFINITION 4.11. *Let X and Y be topological spaces. A sequence (f_n) of functions $f_n : X \to Y$ is said to **converge (pointwise)** to a function $f : X \to Y$ if for each $x \in X$, the sequence $(f_n(x))$ converges in Y to $f(x)$.*

SE 4.19. The sequence (s_n) converges to a function $s : [0, 1] \to [0, 1]$. Determine s, and prove that s is not continuous.

Thus, a sequence of continuous functions does not necessarily converge to a continuous function. However, we are guaranteed that the limit function is continuous if the individual sequences $(f_n(x))$ converge at a uniform rate. This is made explicit for the special case where the range is $\mathbb{R}$, in the following definition:

DEFINITION 4.12. *A sequence of functions* $f_n : X \to \mathbb{R}$ *is said to converge* **uniformly** *to* $f : X \to \mathbb{R}$ *if for every* $\varepsilon > 0$ *there exists* $N \in \mathbb{Z}_+$ *such that* $|f_n(x) - f(x)| < \varepsilon$ *for every* $x \in X$ *and* $n \geq N$.

We can visualize the idea behind uniform convergence by considering the situation graphically. If (f_n) converges uniformly to f, then for every band of vertical thickness 2ε, centered on the graph of f, there exists $N \in \mathbb{Z}_+$ such that for every $n \geq N$, the graph of f_n lies within the band. (See Figure 4.6.)
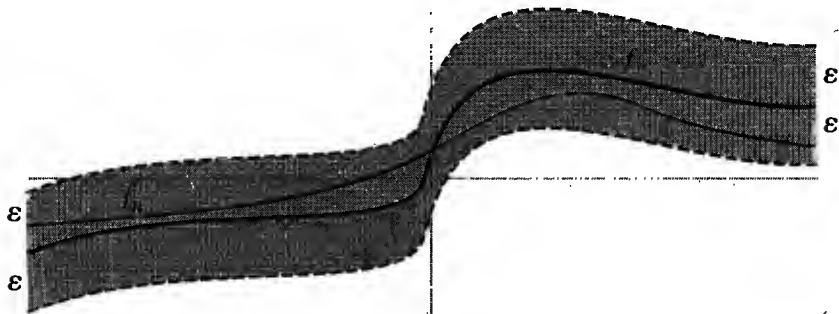
FIGURE 4.6: The graph of f_n lies within ε of the graph of f.

SE 4.20. Prove that the previously defined sequence of functions (s_n) does not converge uniformly.

THEOREM 4.13. *The Uniform Convergence Theorem.* *If* (f_n) *is a sequence of continuous functions* $f_n : X \to \mathbb{R}$ *that converges uniformly to* $f : X \to \mathbb{R}$, *then* f *is continuous.*

> *Proof.* Let $U \subset \mathbb{R}$ be an open set. We prove that for each $x \in f^{-1}(U)$ there exists an open set $V_x \subset X$ such that $x \in V_x \subset f^{-1}(U)$. Let $\varepsilon > 0$ be such that $(f(x^*) - \varepsilon, f(x^*) + \varepsilon) \subset U$. By uniform convergence, we can pick $N \in \mathbb{Z}_+$ such that $|f_n(x) - f(x)| < \frac{\varepsilon}{3}$ for every $n \geq N$ and $x \in X$. Pick $n' \geq N$. Let $U' = (f_{n'}(x) - \frac{\varepsilon}{3}, f_{n'}(x) + \frac{\varepsilon}{3})$, and let $V_x = f_{n'}^{-1}(U')$. We claim that V_x is open in X, contains x, and satisfies $f(V_x) \subset U$. (See Exercise SE 4.21.) Given the claim, it follows that for every $x \in U$, there exists an open set V_x in X such that $x \in V_x \subset f^{-1}(U)$. Therefore U is open in X, and f is continuous. ∎

SE 4.21. In the proof of Theorem 4.13 verify that V_x is open in X, contains x, and satisfies $f(V_x) \subset U$, as claimed.

4.2   Homeomorphisms

In this section we define homeomorphism, a mapping between topological spaces that provides the most fundamental notion of topological equivalence. Homeomorphisms preserve all the properties given by a topology, and thereby define a correspondence between points and between open sets in two topological spaces.

DEFINITION 4.14. *Let X and Y be topological spaces, and let $f : X \to Y$ be a bijection with inverse $f^{-1} : Y \to X$. If both f and f^{-1} are continuous functions, then f is said to be a **homeomorphism**. If there exists a homeomorphism between X and Y, we say that X and Y are **homeomorphic** or **topologically equivalent**, and we denote this by $X \cong Y$.*

Let $f : X \to Y$ be a bijective function. For $f^{-1} : Y \to X$ to be continuous, it must be true that $(f^{-1})^{-1}(U)$ is open in Y for every open set U in X. But since f is a bijection, $(f^{-1})^{-1}(U) = f(U)$ for $U \subset X$. Thus $f(U)$ must be open in Y for every open set U in X. Therefore, saying that f^{-1} is continuous when f is a bijection is equivalent to saying that the image of every open set under f is an open set. Similarly, saying that f is continuous when f is a bijection is equivalent to saying that the image of every open set under the inverse function, f^{-1}, is an open set. (See Figure 4.7.)

We can paraphrase the definition of homeomorphism by saying that f is a homeomorphism if it is a bijection on points and a bijection on the collections of open sets making up the topologies involved. Every point in X is matched to a unique point in Y, with no points in Y left over. At the same time, every open set in X is matched to a unique open set in Y, with no open sets in Y left over. (See Figure 4.7.)
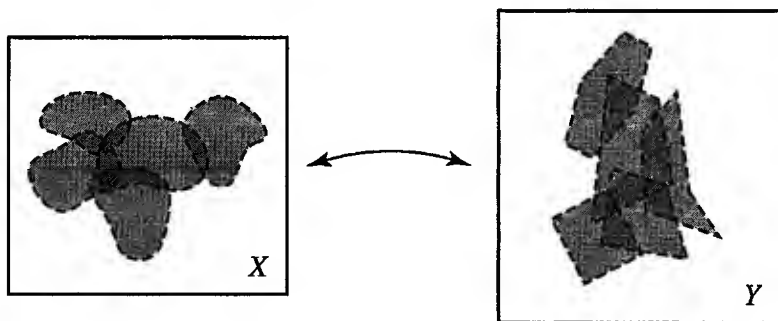
FIGURE 4.7: A homeomorphism sends open sets to open sets.

If you have taken an abstract algebra course, you have likely encountered the concept of a group isomorphism, which is an equivalence between groups. A group isomorphism is a bijection on the elements of the groups that preserves the algebraic structure given by the multiplication operation on each group. In a similar way, a homeomorphism is a bijection on topological spaces that preserves the topological structure given by the collection of open sets in each topological space.

The expression "topologically equivalent," which previously we used informally, has now been made precise by Definition 4.14.

EXAMPLE 4.8. Let X and Y be the topologies on the three-point sets shown in Figure 4.8. Define $f : X \to Y$ by $f(a) = 1$, $f(b) = 2$, $f(c) = 3$. Then f is a homeomorphism since it is a bijection on points and a bijection between the open sets in X and Y.
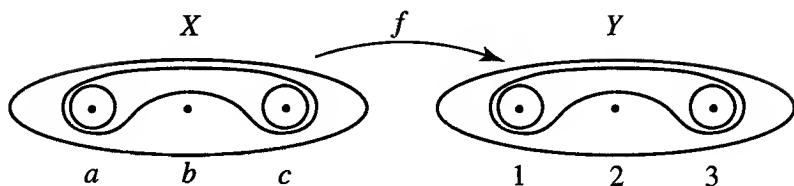
FIGURE 4.8: A homeomorphism from X to Y.

EXAMPLE 4.9. The two topologies on the three-point set $X = \{a, b, c\}$, shown in Figure 4.9, are different as collections of subsets of X but are topologically equivalent since the function $f : X \to X$, given by $f(a) = c$, $f(b) = b$, $f(c) = a$, is a homeomorphism.
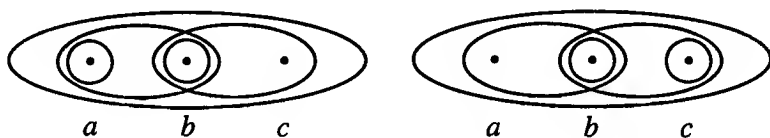
FIGURE 4.9: Two homeomorphic topologies on $X = \{a, b, c\}$.

EXAMPLE 4.10. Define the function $f : \mathbb{R} \to \mathbb{R}$ by $f(x) = 3x + 1$. This function is a bijection with inverse given by $f^{-1}(x) = \frac{x-1}{3}$. (See Example 0.12.) Since f and f^{-1} are both polynomials on $\mathbb{R}$, they are both continuous by Theorem 4.4. Thus, f is a homeomorphism.

Since f is a homeomorphism, it preserves the topological properties of $\mathbb{R}$. Stretching by a factor of three and then sliding to the right by a distance of one does not change the basic topological properties of the real line.

The following basic facts about homeomorphisms together imply that topological equivalence is an equivalence relation on the collection of all topological spaces. (See Exercise 4.28.)

(i) The function $id : X \to X$, defined by $id(x) = x$, is a homeomorphism.

(ii) If $f : X \to Y$ is a homeomorphism, then so is $f^{-1} : Y \to X$.

(iii) If $f : X \to Y$ and $g : Y \to Z$ are homeomorphisms, then so is $g \circ f : X \to Z$.

EXAMPLE 4.11. Let $(-1, 1)$ have the standard topology, and define a function $f : \mathbb{R} \to (-1, 1)$ by $f(x) = \frac{x}{1+|x|}$. The graph of f is shown in Figure 4.10. It is apparent from the graph that f is a bijection between $\mathbb{R}$ and $(-1, 1)$. It is also apparent that preimages of open intervals are open intervals for both f and f^{-1}. Therefore, f and f^{-1} are both continuous, and it follows that f is a homeomorphism between $\mathbb{R}$ and $(-1, 1)$.

$$f(x) = \frac{x}{1 + |x|} \qquad (-1, 1)$$
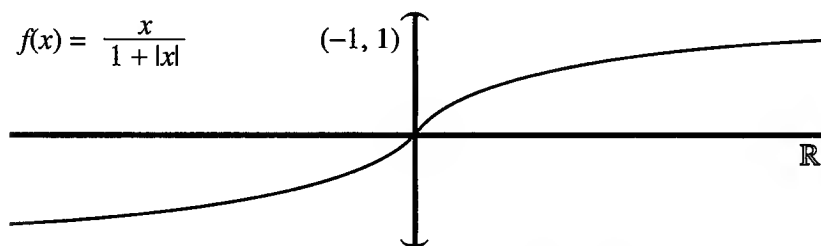
FIGURE 4.10: A homeomorphism between $\mathbb{R}$ and $(-1, 1)$.

EXAMPLE 4.12. Not only is $(-1, 1)$ homeomorphic to $\mathbb{R}$, but every nonempty open interval (a, b) is as well. In fact, consider the following collections of intervals with the standard topology (assuming a and b are arbitrary real numbers with $a < b$):

(i) Open intervals: (a, b), $(-\infty, a)$, (a, ∞), $\mathbb{R}$,

(ii) Closed, bounded intervals: $[a, b]$,

(iii) Half-open intervals and closed, unbounded intervals:

$$[a, b), (a, b], (-\infty, a], [a, \infty).$$

Within each of the collections (i), (ii), and (iii), all of the spaces are topologically equivalent. We illustrate example homeomorphisms in Figure 4.11 and discuss these equivalences further here.

The function $f : \mathbb{R} \to (a, \infty)$, given by $f(x) = e^x + a$, is a homeomorphism. Thus, $\mathbb{R}$ is homeomorphic to every interval (a, ∞). Since topological
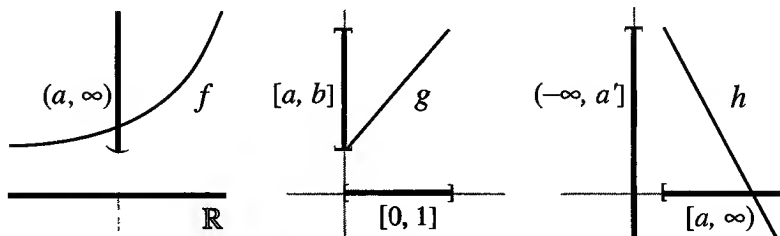
FIGURE 4.11: The homeomorphisms f, g, and h.

equivalence is an equivalence relation, it also follows that every interval (a, ∞) is homeomorphic to every other interval in the form (a', ∞).

The linear function $g : [0, 1] \to [a, b]$, given by $g(x) = (b - a)x + a$, is a homeomorphism between $[0, 1]$ and $[a, b]$. Therefore, every interval $[a, b]$ is homeomorphic to $[0, 1]$, and, consequently, every interval $[a, b]$ is homeomorphic to every other closed, bounded interval $[a', b']$ with $a' < b'$.

Furthermore, the function $h : [a, \infty) \to (-\infty, a']$, given by $h(x) = -x + a' + a$, is a homeomorphism between the intervals $[a, \infty)$ and $(-\infty, a']$. Thus, if I_1 and I_2 are intervals of either form, $[a, \infty)$ or $(-\infty, a']$, then I_1 and I_2 are homeomorphic.

In Exercises 4.25 and 4.26 you are asked to define homeomorphisms that establish the remainder of the topological equivalences in the claim that all of the interval subspaces in each collection are topologically equivalent.

From one collection to another, the interval subspaces of $\mathbb{R}$ are not homeomorphic. Although that might seem intuitively clear, we need to develop further topological concepts before we can explicitly prove this. We do so in Chapter 6. (See Exercise 6.27.)

EXAMPLE 4.13. Let $[0, 2\pi)$ and S^1 have the standard topology as subspaces of $\mathbb{R}$ and $\mathbb{R}^2$, respectively. We denote each point in S^1 by p_θ, where p_θ represents the point on S^1 at angle $\theta \in \mathbb{R}$, measured counterclockwise from the positive x-axis. Define $f : [0, 2\pi) \to S^1$ by $f(\theta) = p_\theta$, as illustrated in Figure 4.12. It is clear that f is a bijection.
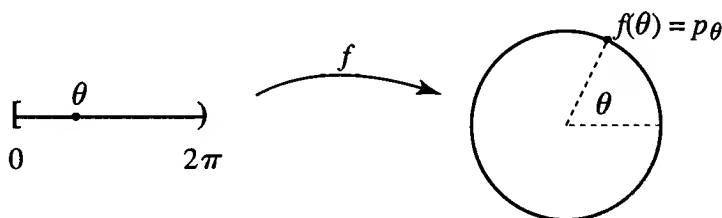
FIGURE 4.12: A continuous bijection from $[0, 2\pi)$ onto S^1 that is not a homeomorphism.

Is f continuous? Yes. We can see this by looking at the preimage of an arbitrary basis element for the topology on S^1. Such a basis element can be expressed in the form $B = \{p_\theta \in S^1 \mid a < \theta < b \text{ and } a, b \in \mathbb{R}\}$. We consider two cases for B: either p_0 is in B or it is not, where p_0 is the point corresponding to angle $\theta = 0$. If $p_0 \notin B$, then $f^{-1}(B)$ is an open interval (c, d) in $[0, 2\pi)$. If $p_0 \in B$, then $f^{-1}(B)$ is in the form $[0, e) \cup (g, 2\pi)$. In either case, $f^{-1}(B)$ is an open subset of $[0, 2\pi)$ in the standard topology. It follows that f is continuous.

But is f^{-1} continuous? No! The half-open interval $[0, 1/4)$ is open in the domain $[0, 2\pi)$, but $f([0, 1/4))$ is not open in S^1. Therefore f is not a homeomorphism.

In fact, there are no homeomorphisms between these two spaces, and therefore they are not topologically equivalent. We will see how to show this when we establish some properties involving connectedness, in Chapter 6. (See Exercise 6.28.)

EXAMPLE 4.14. In the standard topology on each, the plane is topologically equivalent to the open right half plane $H = \{(x, y) \in \mathbb{R}^2 \mid x > 0\}$ and the open disk $\mathring{D} = \{(x, y) \in \mathbb{R}^2 \mid x^2 + y^2 < 1\}$, as illustrated in Figure 4.13.
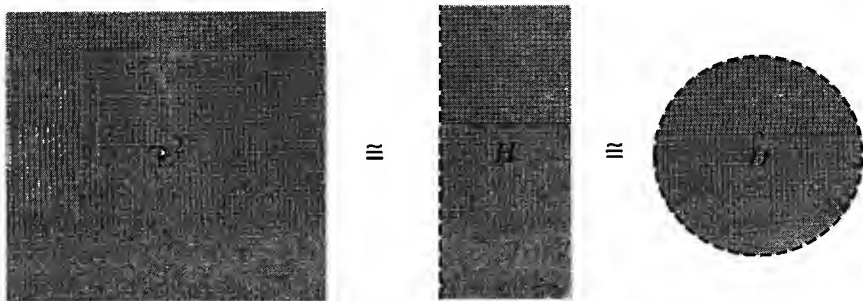
FIGURE 4.13: The plane is homeomorphic to the open half plane and the open disk.

The function $f : \mathbb{R}^2 \to H$, defined by $f(x, y) = (e^x, y)$, is a homeomorphism between $\mathbb{R}^2$ and H. It maps $\mathbb{R}^2$ to H, sending vertical lines to vertical lines, as follows:

    (i) The left half plane is mapped to the strip in H where $0 < x < 1$.

    (ii) The y-axis is mapped to the line $x = 1$.

    (iii) The right half plane is mapped to the region in H where $x > 1$.

The function $g : \mathbb{R}^2 \to \mathring{D}$, defined (in polar coordinates) by $g(r, \theta) = (\frac{r}{1+r}, \theta)$, is a homeomorphism between $\mathbb{R}^2$ and $\mathring{D}$. It contracts the whole plane radially inward to coincide with the open disk $\mathring{D}$.

EXAMPLE 4.15. The surface of a cube C is homeomorphic to the sphere S^2, as illustrated in Figure 4.14. If we regard each as centered at the origin in 3-space, then the function $f : C \to S^2$, defined by $f(p) = p/|p|$, is a homeomorphism. As we see in the figure, f maps points in C bijectively to points in S^2 and maps the collection of open sets in C bijectively to the collection of open sets in S^2.
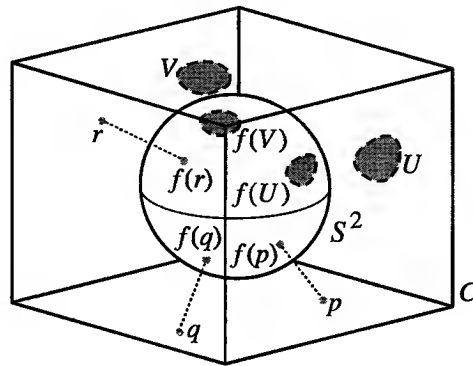
FIGURE 4.14: The surface of a cube is homeomorphic to the sphere.

In the next example, we see that removing a point from a sphere yields a space that is homeomorphic to the plane.

EXAMPLE 4.16. In $\mathbb{R}^3$, let $N = (0, 0, 1)$ be the point at the north pole in the sphere S^2. Let $X = S^2 - \{N\}$, and give X the subspace topology it inherits from $\mathbb{R}^3$. The space X is called a **punctured sphere**. We show that X is homeomorphic to the plane $\mathbb{R}^2$. Note that the xy-plane, as a subspace of $\mathbb{R}^3$, is naturally homeomorphic to $\mathbb{R}^2$, so we view $\mathbb{R}^2$ as that subspace.

We define a function $f : X \to \mathbb{R}^2$ as follows. For each point $p = (x, y, z)$ in X, take the ray that begins at N and passes through p. We define $f(p)$ to be the unique point where the ray passes through the xy-plane. (See Figure 4.15.) The map f is called a **stereographic projection**. In Exercise 4.27 we ask you to derive a specific formula for f.
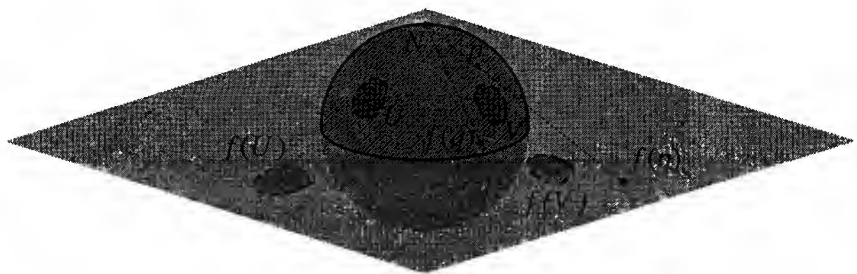
FIGURE 4.15. Stereographic projection.

It is clear from the illustration that f is a bijection and that f maps the collection of open sets in X bijectively to the collection of open sets in the plane. Hence, f is a homeomorphism from X to the plane. Thus, removing a point from the sphere results in a space that is topologically equivalent to the plane.

In the next example we see how collapsing each end in a cylinder results in a space that is homeomorphic to the sphere.

EXAMPLE 4.17. Let X be the cylinder $S^1 \times [-1, 1]$ in $\mathbb{R}^3$. We create a quotient space X^* by defining a partition of X that is made up of the following sets:

   (i) Each one-point set $\{(x, y, z)\}$, where $z \in (-1, 1)$ and $(x, y) \in S^1$,

   (ii) The set $T = \{(x, y, 1) \mid (x, y) \in S^1\}$ at the top of the cylinder,

   (iii) The set $B = \{(x, y, -1) \mid (x, y) \in S^1\}$, at the bottom of the cylinder.

A quotient topology is defined on X^* by the function $p : X \to X^*$ that takes each point in X and maps it to the set in X^* containing it.

In Figure 4.16 we illustrate X^* and two example open sets in X^*. In the identification from X to X^*, the sets T and B are collapsed to individual points, T^* and B^*, respectively. An open set U in X^* that does not contain T^* or B^* is essentially the same set as its corresponding open set $p^{-1}(U)$ in X. An open set V in X^* that contains the point T^*, but not the point B^*, corresponds to an open set $p^{-1}(V)$ in X that contains the subset T and is disjoint from B. We can determine other open sets in X^* in a similar manner.
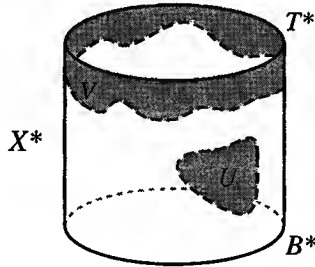
FIGURE 4.16: The sets U and V are open sets in X^*.

We show that the quotient space X^* is homeomorphic to the sphere S^2. In 3-space, the cylinder surrounds the sphere, and both have the z-axis as a central vertical axis, as shown in Figure 4.17. Furthermore, the cylinder and sphere intersect in the circle of radius 1, centered at the origin in the xy-plane, and the north (N) and south (S) poles of the sphere are at the center of the top (T) and bottom (B) circles of the cylinder, respectively.

To obtain a homeomorphism $g : X^* \to S^2$, we first define a function $f : X \to S^2$ and then use f to define g. The function f is defined by projecting each point p in the cylinder $X = S^1 \times [-1, 1]$ along a horizontal line, directly toward the z-axis, onto the point $f(p)$ on the sphere. We can express this function in cylindrical coordinates in $\mathbb{R}^3$ as $f(1, \theta, z) = (\sqrt{1 - z^2}, \theta, z)$. The function f maps the cylinder onto the sphere with the following properties, as illustrated in Figure 4.17:

   (i) The open subset of the cylinder, lying between the top and bottom circles, is mapped bijectively to the sphere minus its north and south poles.

(ii) The entire top circle of the cylinder, the set T, is mapped to the north pole, N.

(iii) The entire bottom circle of the cylinder, the set B, is mapped to the south pole, S.
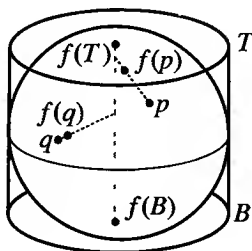
FIGURE 4.17: The cylinder is mapped onto the sphere by f.

Since f is constant on each of the circles T and B, and since each of these sets is considered a single element in X^*, it follows that f induces a bijective function $g : X^* \to S^2$.

To show that g is a homeomorphism, we check that g and g^{-1} map open sets to open sets. Let U be an open set in X^*, and suppose it does not contain either T^* or B^* as points. Then $g(U)$ is an open set in S^2 that is disjoint from both N and S. (See Figure 4.18.) However, if V is an open set in X^* that contains T^*, then g maps V to an open set in S^2 that contains N. The other cases for open sets in X^* are similar. Also, the same type of argument can be used to show that g^{-1} maps open sets in S^2 to open sets in X^*. It follows that g is a homeomorphism between X^* and S^2.
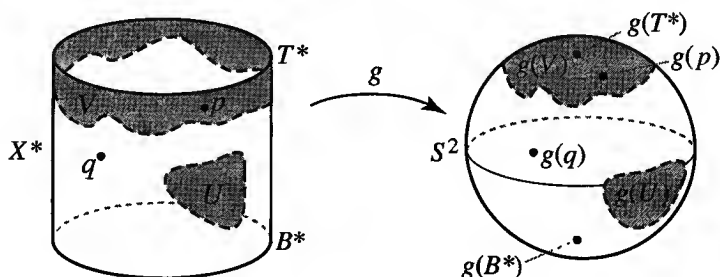
FIGURE 4.18: The quotient space X^* is homeomorphic to the sphere S^2.

EXAMPLE 4.18. In Figure 4.19 we see two bands in 3-space, each with the standard topology. Because we cannot perform a rubber-sheet deformation from one to the other, they might seem to be topologically distinct, but in fact they are homeomorphic. We define a map from the first band to the second by cutting the first band open along segment S, untwisting the resulting strip, and then gluing the cut segments back together so that they match as they did

beforehand. The map takes a point on the first band, traces it through this process, and maps it to the point where it ends up on the second band. This map is a bijection on the bands and on the collections of open sets in the bands, and therefore is a homeomorphism.
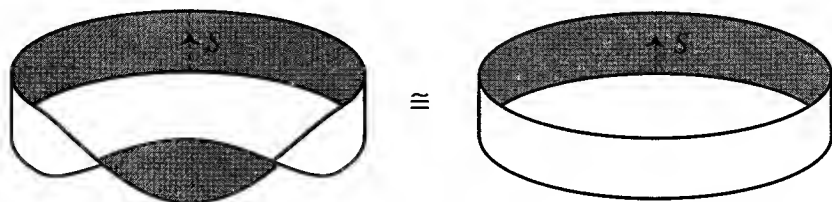
FIGURE 4.19: Two homeomorphic bands in 3-space.

The band on the left in Figure 4.19 is not a Möbius band. It is made by putting a full twist in a strip before gluing the ends together, while a Möbius band is made by putting a half twist in a strip before gluing the ends together. (See Example 3.19.)

Example 4.18 raises an important point about equivalence in topology. If we cut an object, we might be left with an object that is not topologically equivalent to the original. But, if we glue the cut back together properly, with points along either side of the cut matching as they did before the cut, then no matter how much twisting and deforming we do in the meantime, the result is homeomorphic to the original.

Although homeomorphism represents the most fundamental equivalence between two topological spaces, it is not the equivalence most people envision when they think of topology as rubber-sheet geometry. As already indicated, the two bands appearing in Figure 4.19 cannot be deformed from one to the other as rubber sheets, but they are homeomorphic. The rubber-sheet notion of equivalence is made precise in Chapter 12 via what is known as isotopy, a continuous, parameterized family of homeomorphisms.

There is a distinct difference in how the bands in Example 4.18 sit in 3-space. We can see this difference by examining the bands' edges, the curves E_1 and E_2 shown in Figure 4.20. On the left side of the figure, the edges are linked, while on the right side they are not. This idea will be formalized in the concept of linking number, which we introduce in Chapter 12.
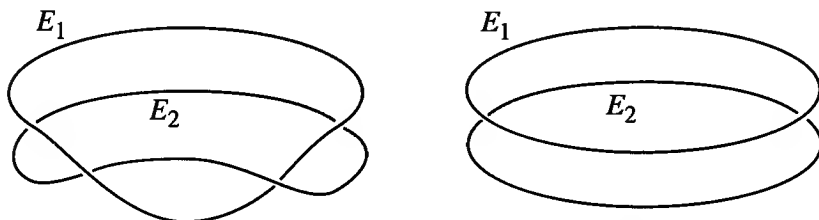
FIGURE 4.20: The edges E_1 and E_2 are linked on the left, but not on the right.

The idea of one topological space sitting inside another is captured by the following definition:

DEFINITION 4.15. *An **embedding** of X in Y is a function $f : X \to Y$ that maps X homeomorphically to the subspace $f(X)$ in Y.*

We think of an embedding $f : X \to Y$ as placing a copy of X in Y, as shown in Figure 4.21.
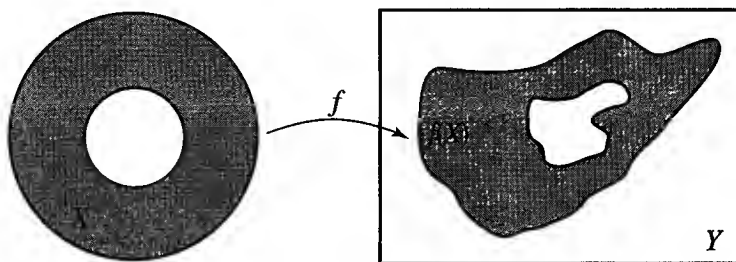
FIGURE 4.21: Embedding X in Y.

Each of the bands appearing in Figure 4.19 can be regarded as the image of an embedding of an annulus in 3-space. While the bands are homeomorphic as subspaces of $\mathbb{R}^3$, we will see that the corresponding embeddings are distinct once we define a notion of equivalence between embeddings, in Chapter 12.

DEFINITION 4.16. *Let X be a topological space. If $f : [-1, 1] \to X$ is an embedding, then the image of f is called an **arc** in X, and if $f : S^1 \to X$ is an embedding, then the image of f is called a **simple closed curve** in X. (See Figure 4.22.)*
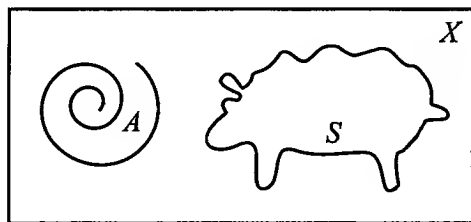
FIGURE 4.22: An arc A and a simple closed curve S in X.

In Figure 4.23, we depict some simple closed curves in 3-space. Embeddings $f : S^1 \to \mathbb{R}^3$, naturally enough, are referred to as knots and are the subject of the study of knot theory. We discuss embeddings further in Chapter 11 and knot theory in Chapter 12.

Certain properties defined for topological spaces are preserved by homeomorphisms. An example is provided by the next theorem.
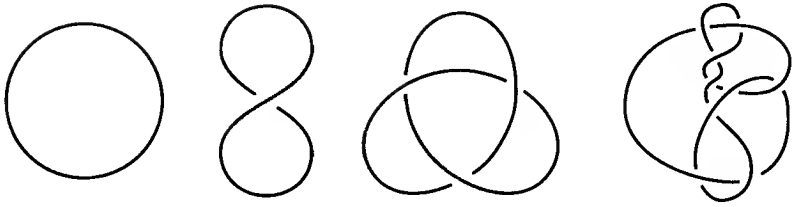
FIGURE 4.23: Knots are embeddings of the circle in 3-space.

THEOREM 4.17. *If $f : X \to Y$ is a homeomorphism and X is Hausdorff, then Y is Hausdorff.*

> ***Proof.*** Suppose that X is Hausdorff and $f : X \to Y$ is a homeomorphism. Let x and y be distinct points in Y. Then $f^{-1}(x)$ and $f^{-1}(y)$ are distinct points in X. Thus, there exist disjoint open sets U and V containing $f^{-1}(x)$ and $f^{-1}(y)$, respectively. It follows that $f(U)$ and $f(V)$ are disjoint open sets containing x and y, respectively. Therefore Y is Hausdorff. ∎

Theorem 4.17 implies that $\mathbb{R}$ in the standard topology is not homeomorphic to $\mathbb{R}$ in the finite complement topology because the former is Hausdorff while the latter is not.

A property of topological spaces that is preserved by homeomorphism is said to be a **topological property**. Theorem 4.17 implies that being Hausdorff is a topological property. In general, properties that are defined in terms of open sets, like Hausdorff, are topological properties.

As we have seen in a few examples in this section, we can prove that two topological spaces are homeomorphic by defining a homeomorphism between them. Conversely, in order to prove that two topological spaces are not homeomorphic, it is necessary to show that none of the functions defined between them is a homeomorphism. However, it is often too difficult to consider every function between the spaces. Instead, as in our argument that $\mathbb{R}$ in the standard topology is not homeomorphic to $\mathbb{R}$ in the finite complement topology, we find a topological property held by one space but not the other. Then we know that the two spaces are not homeomorphic.

Exercises for Section 4.2

4.22. Consider all of the possible topologies on the two-point set $X = \{a, b\}$. Indicate which ones are homeomorphic.

4.23. Find three different topologies on the three-point set $X = \{a, b, c\}$, each consisting of five open sets (including X and $\varnothing$), such that two of the topologies are homeomorphic to each other, but the third is not homeomorphic to the other two.

4.24. Prove that a bijection $f : X \to Y$ is a homeomorphism if and only if f and f^{-1} map closed sets to closed sets.

4.25. (a) Provide a formula for a homeomorphism between $\mathbb{R}$ and the interval $(-\infty, a)$.

(b) Provide a formula for a homeomorphism between $\mathbb{R}$ and the interval (a, b), with $a < b$.

(c) Given the homeomorphisms in Example 4.12 and the first two parts of this exercise, prove that if I_1 and I_2 are intervals in collection (i) in Example 4.12, then I_1 and I_2 are topologically equivalent.

4.26. (a) Provide a formula for a homeomorphism between the intervals $[0, \infty)$ and $[a, b)$, with $a < b$.

(b) Provide a formula for a homeomorphism between the intervals $(-\infty, 0]$ and $(a, b]$, with $a < b$.

(c) Given the homeomorphisms in Example 4.12 and the first two parts of this exercise, prove that if I_1 and I_2 are intervals in collection (iii) in Example 4.12, then I_1 and I_2 are topologically equivalent.

4.27. Provide an explicit formula for the stereographic projection function in Example 4.16.

4.28. Prove each of the following statements, and then use them to show that topological equivalence is an equivalence relation on the collection of all topological spaces:

(a) The function $id : X \to X$, defined by $id(x) = x$, is a homeomorphism.

(b) If $f : X \to Y$ is a homeomorphism, then so is $f^{-1} : Y \to X$.

(c) If $f : X \to Y$ and $g : Y \to Z$ are homeomorphisms, then so is the composition $g \circ f : X \to Z$.

4.29. Show that $\mathbb{R}^2 - \{O\}$ in the standard topology is homeomorphic to $S^1 \times \mathbb{R}$.

4.30. Find two distinct topologies on $\mathbb{R}$ such that the first is strictly finer than the second but the two of them are homeomorphic with each other.

4.31. (a) Show that having nonempty open sets that contain finitely many points is a topological property.

(b) Prove that the digital line is not homeomorphic to $\mathbb{Z}$ with the finite complement topology.

4.32. Show that homeomorphism preserves interior, closure, and boundary as indicated in the following implications:

(a) If $f : X \to Y$ is a homeomorphism, then $f(\text{Int}(A)) = \text{Int}(f(A))$ for every $A \subset X$.

(b) If $f : X \to Y$ is a homeomorphism, then $f(\text{Cl}(A)) = \text{Cl}(f(A))$ for every $A \subset X$.

(c) If $f : X \to Y$ is a homeomorphism, then $f(\partial(A)) = \partial(f(A))$ for every $A \subset X$.

4.33. Let $X \times Y$ be partitioned into subsets of the form $X \times \{y\}$ for all y in Y. If we let $(X \times Y)^*$ denote the collection of sets in the partition, show that $(X \times Y)^*$ with the resulting quotient topology is homeomorphic to Y.

4.34. Let X, Y, and Z be topological spaces. Prove that the three product spaces $(X \times Y) \times Z$, $X \times (Y \times Z)$, and $X \times Y \times Z$ are homeomorphic to each other.

4.35. In Chapter 3 we introduced three different representations of the torus:

(i) The definition of the torus as a subspace of $\mathbb{R}^3$ in Example 3.5,

(ii) The product space $S^1 \times S^1$ in Example 3.10, and

(iii) A quotient space obtained by gluing edges of a square in Example 3.20.

Verify that these spaces are topologically equivalent to each other by defining homeomorphisms between them.

4.36. In this exercise we establish that products of homeomorphic spaces are homeomorphic. Let $f : X \to Y$ and $g : X' \to Y'$ be homeomorphisms. Prove that the function $h : X \times X' \to Y \times Y'$, defined by $h(x, x') = (f(x), g(x'))$, is a homeomorphism.

4.37. Begin with a collection of topological graphs representing the letters of the alphabet. (See Exercise 3.34.) By visual inspection, determine which topological graphs in the collection are homeomorphic and which are not, then group them into equivalence classes by topological equivalence. (Note: In Chapter 6 we introduce properties that enable us to verify that topological graphs from the different equivalence classes in this exercise are not homeomorphic—see Exercise 6.31.)

4.3   The Forward Kinematics Map in Robotics

In the field of robotics, there is a naturally defined continuous function, called the forward kinematics map, which plays a role in motion planning for linkages, robot arms, and other similar mechanisms. In this section we present an introduction to the forward kinematics map by considering a few examples and examining some interesting properties and problems involving this function.

Recall the linkage presented in Example 3.29, where we had two rods, pictured again in Figure 4.24. We assume throughout this section that rod B is shorter than rod A. The configuration space for the system is the torus, $S^1 \times S^1$, where the first S^1 corresponds to the circle of angles θ_A through which rod A can turn about its fixed end, and the second S^1 corresponds to the circle of angles θ_B through which rod B can turn about its end that is fixed to rod A.
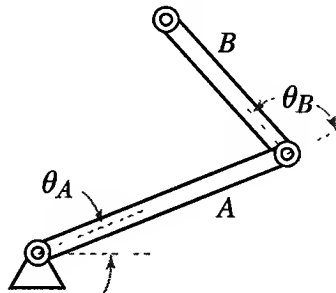
FIGURE 4.24: A two-rod linkage.

The operational space for the linkage is the space traced out by the end of rod B. In this particular situation, we can picture the linkage as a drawing device, and the operational space as the set of all points we can color with a pen at the end of rod B. The resulting operational space is an annulus. (See Figure 3.40.)

In general, in studying the design of a machine or robot arm, we are interested in a particular point on the mechanism where there is a tool that serves a specific function, such as spraying paint, picking up a part, drilling a hole, and so on. This point is called the **end effector**, and the operational space is the space traced out by the end effector. For our purposes, it suffices to view the end effector as a point where there is a pen on a drawing device.

To each point in the configuration space of a mechanism, we associate the corresponding end-effector point in the operational space. Thus, a function f is defined; it is called the **forward kinematics map** for the mechanism. It is natural to assume that the forward kinematics map is continuous because points close together in the configuration space correspond to points close together in the operational space. In the case of our two-rod linkage, f is a continuous map from the torus to the annulus, as shown in Figure 4.25.
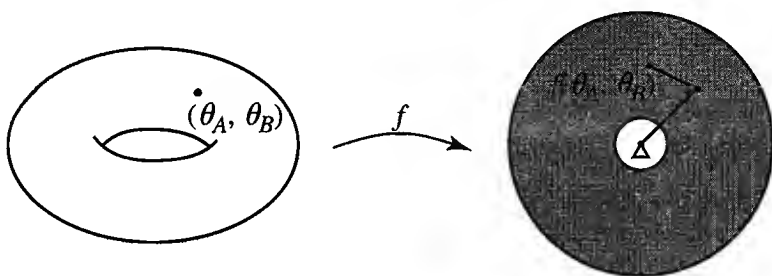
FIGURE 4.25: The forward kinematics map for the two-rod linkage.

An important question in the field of robotics is whether a given path in the operational space of a mechanism can be traced by the end effector. In our example, we can ask if there is a way to manipulate the configuration space variables θ_A and θ_B to yield a given path in the operational space annulus. (See Figure 4.26.) In other words, can the linkage draw a given curve in the annulus?

FIGURE 4.26: Can the linkage trace out the given path?

Let us make this idea more mathematically precise.

DEFINITION 4.18. *Let Y be a topological space, and let $[0, 1] \subset \mathbb{R}$ have the standard topology. A **path** in Y is a continuous function $p : [0, 1] \to Y$. We say that the path **begins at** $p(0)$ and **ends at** $p(1)$. (See Figure 4.27.)*
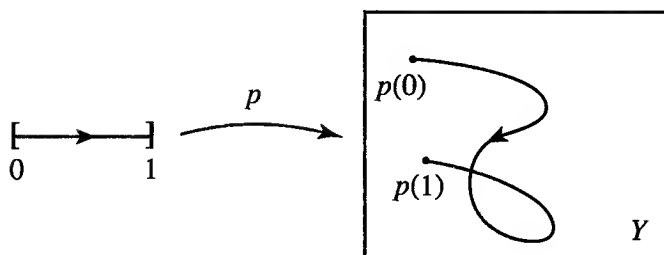
FIGURE 4.27: A path in Y.

We can think of a path as a trajectory in Y, traced from time 0 to time 1. While a path is a function, we often use the term path to refer to the image of a path.

Suppose that we have a linkage with configuration space C, operational space O, and forward kinematics map $f : C \to O$. Given an operational space path $p : [0, 1] \to O$, our path-tracing question now asks: Is there a configuration space path $q : [0, 1] \to C$ such that $p = f \circ q$? (See Figure 4.28.) If such a q exists, we say that q **traces** p.
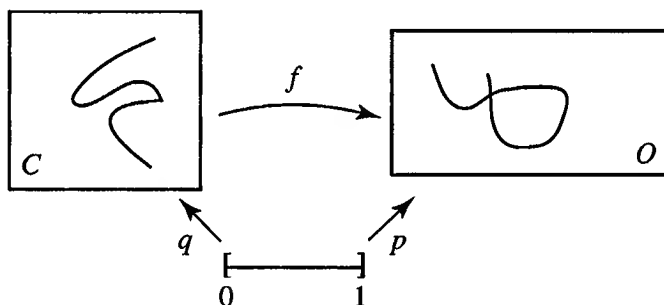
FIGURE 4.28: The path q traces path p if $p = f \circ q$.

To say that q traces p means that there is a way to move the linkage, specified by the path q in the configuration space, to obtain the desired path p in the operational space. The problem of finding a configuration space path corresponding to a given operational space path is known as an **inverse kinematics problem**.

EXAMPLE 4.19. Consider the two-rod linkage shown in Figure 4.24. We represent the configuration space torus as a square with its edges identified as in Section 3.3. The horizontal direction on the square corresponds to angle θ_A, and the vertical direction corresponds to θ_B.

Let p be a circular path in the operational space, as shown in Figure 4.29. Then p can be traced by a configuration space path q with θ_A fixed at θ^* and θ_B running from 0 to 2π.
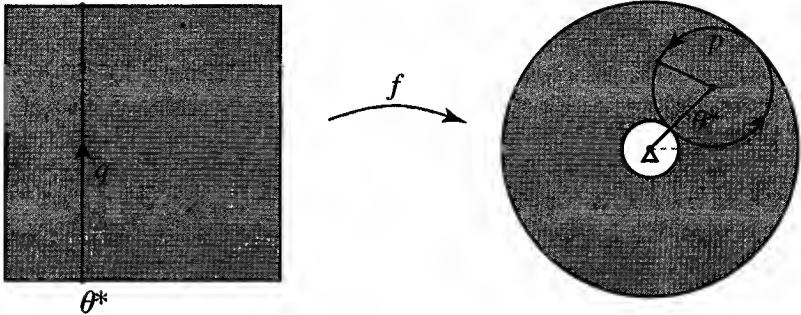
FIGURE 4.29: Configuration space path q traces out operational space path p.

Now let p be a circular path around the center of the operational space annulus, as shown in Figure 4.30. In this case, p is traced by a configuration space path q with θ_B fixed at θ^* and θ_A running from 0 to 2π.
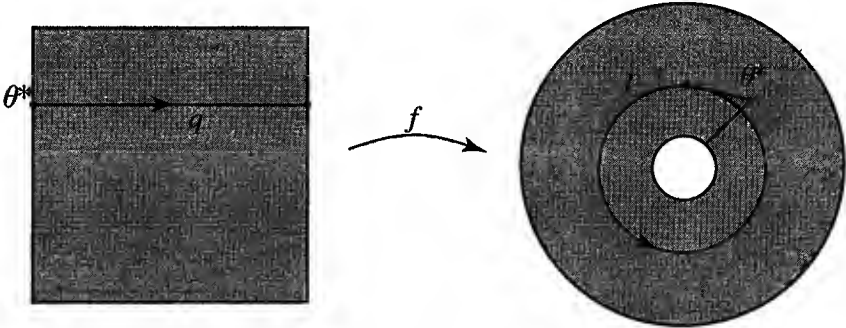
FIGURE 4.30: Configuration space path q traces operational space path p.

Consider the two operational space paths shown in Figure 4.31. Path p_1 is a circular path that is tangent to the inner and outer circles of the annulus. Path p_2 follows two circular arcs that have a radius equal to the length of rod B. In Exercise 4.38 we ask you to find configuration space paths that trace these operational space paths.
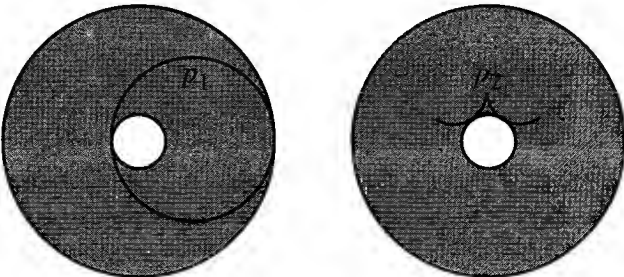
FIGURE 4.31: Are there configuration space paths that trace the paths shown?

While it might be possible to find a configuration space path that enables a mechanism to follow a specified path in the operational space, sometimes the process results in problematic configurations of the mechanism. In the next example, we see such a situation and how it impacts spaceflight navigational systems.

EXAMPLE 4.20. Consider the mechanism illustrated in Figure 4.32. Suppose that the components of the mechanism can rotate as indicated in the figure. The rotating components are called gimbals, and since each one can rotate through a full circle, the configuration space of the mechanism is $S^1 \times S^1 \times S^1$, the 3-torus.
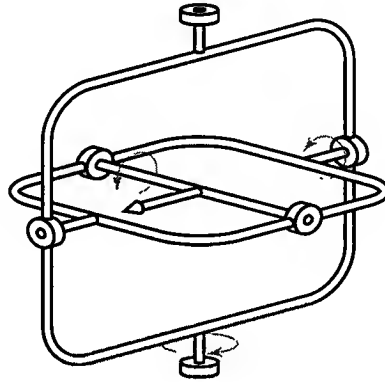
FIGURE 4.32: A three-gimbal mechanism.

The operational space is the space traced out by the tip of the pointer on the inner gimbal. The other end of the pointer lies on all three axes of rotation and therefore is fixed. Furthermore, it is easy to see that the tip of the pointer can be oriented in any direction in 3-space, and therefore the operational space is a sphere.

While the mechanism allows the arrow to point in any direction, there are configurations for which freedom of motion is restricted. These configurations of the mechanism give rise to the phenomenon known as **gimbal lock** and occur when the three gimbals are coplanar, as shown in Figure 4.33. With the gimbals thus configured, rotation of either the outer gimbal or the inner gimbal results in the arrow tracing a path on the equator of the operational space sphere. In contrast, rotation of the middle gimbal keeps the arrow fixed on the same point x. It follows that every smooth motion of the mechanism through this configuration results in a path that is tangent to the pictured equator at x. Consequently, if the mechanism passes through this configuration and generates a path that is not tangent to the equator, then the mechanism must come to rest at this configuration. The configuration space point associated with such a configuration is called a **singular point** of the forward kinematics map.

We do not provide a formal definition of singular points of the forward kinematics map since it requires the development of concepts from the fields
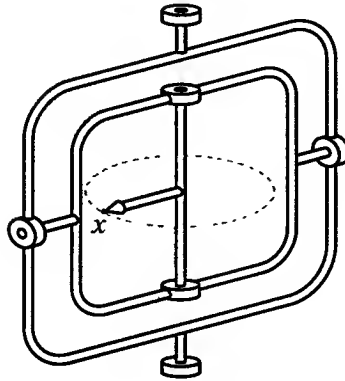
FIGURE 4.33: The gimbal-lock configuration.

of differential geometry and differential topology. It is important to realize, though, that singular points are points in configuration space corresponding to configurations of the mechanism, rather than points in the operational space. In this example, if the gimbals are configured as in Figure 4.32, then the corresponding point in operational space is the same point x as in the gimbal-lock configuration depicted in Figure 4.33. The configuration in Figure 4.32 is not singular and poses no restrictions on tracing paths through x.

The configuration associated with gimbal lock posed a navigational challenge during the National Aeronautics and Space Administration's manned spaceflights to the moon in the late 1960s and early 1970s. Each spacecraft's navigational system included a gimballed mechanism similar to the one shown in Figure 4.32. We present a simplified overview of that system, but the problems caused by singular points in the configuration space are essentially the same.

The navigational system's inertial measurement unit (IMU) was attached to the spacecraft and contained a system of three gimbals in the arrangement shown in Figure 4.34. The inner gimbal held the stable platform, the main functional component of the IMU. Within the stable platform there were electrical and mechanical systems that served two important purposes. First, they
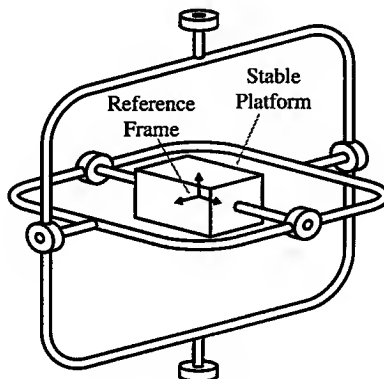
FIGURE 4.34: The structure of the inertial measurement unit (IMU).

defined a reference frame for navigation in space. Second, as long as gimbal lock was avoided, they kept the stable platform in a fixed orientation when the spacecraft maneuvered, thereby keeping the reference frame constant.

Gimbal lock occurs when the gimballed device is in the singular position shown in Figure 4.35. In this position, if the spacecraft turns on an axis perpendicular to the plane of the gimbals, then the three gimbals turn together in their plane, maintaining the gimbal lock configuration and rotating the stable platform so that the reference frame is lost. In fact, at gimbal lock every spacecraft motion including such a rotational component causes the reference frame to be lost.
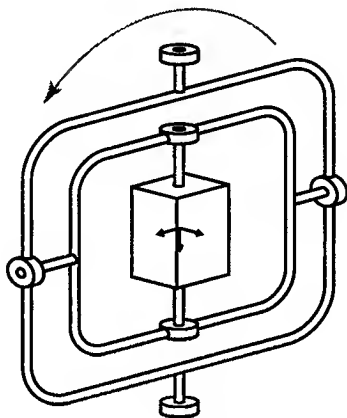
FIGURE 4.35: The reference frame can be lost as a result of a rotation in the gimbal-lock configuration.

Gimbal lock was such a concern on these spaceflights, that a warning indicator was used to alert the astronauts when gimbal lock was imminent. The astronauts would then maneuver the spacecraft to enable the IMU to move away from gimbal lock, thereby averting the time-intensive task of resetting the reference frame.

As is evident in the examples in this section, an analysis of the forward kinematics map is important in the design of mechanical systems. The forward kinematics map can provide critical information about impossible, restrictive, or potentially problematic configurations of a mechanism under consideration.

Exercises for Section 4.3

4.38. Consider the two-rod linkage in Figure 4.24 and the two operational space paths shown in Figure 4.31. In diagrams of the configuration space torus, sketch paths q_1 and q_2 that trace paths p_1 and p_2, respectively.

4.39. Consider the two-rod linkage in Figure 4.24. A configuration (θ_A, θ_B) of this linkage is a singular configuration if at configuration (θ_A, θ_B) the operational space path obtained by holding θ_A fixed and varying θ_B is tangent to the path obtained by holding θ_B fixed and varying θ_A. Determine the singular configurations of this linkage.

4.40. Consider the linkage, called Watt's Parallel Motion, illustrated in Figure 4.36(i). Each rod A and B has one end pinned in the plane and the other end connected to an end of rod C. The rods can rotate at each pinned point and at each connection point. Imagine we have a pen at the midpoint of rod C. As illustrated, θ_A is the angle that rod A makes, measured counterclockwise from the horizontal axis; the same holds for θ_B and B. The configuration space is the set of pairs of angles (θ_A, θ_B) corresponding to possible configurations of the linkage; it is illustrated as a subspace of the $\theta_A\theta_B$-plane in Figure 4.36(ii). The operational space is the subspace of the plane traced out by the pen; it is illustrated in Figure 4.36(iii). You might want to make a working model of the linkage to see how it operates.
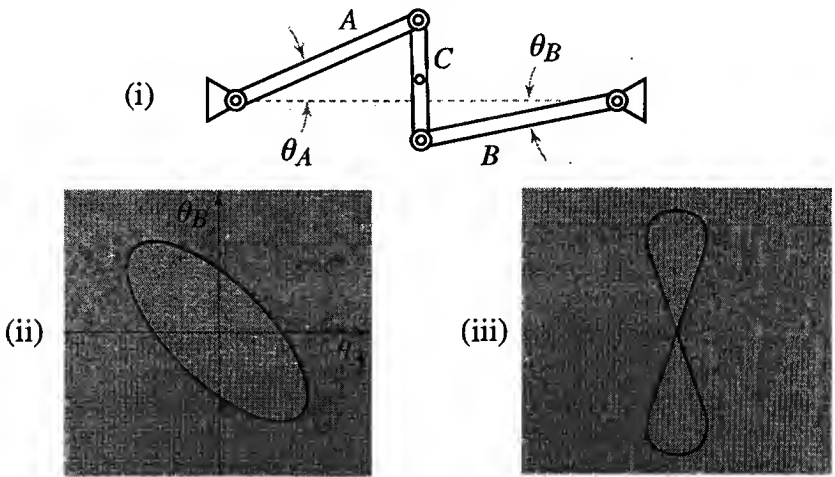
FIGURE 4.36: Watt's Parallel Motion, its configuration space, and its operational space.

(a) Consider the set of configuration space points (θ_A, θ_B). Explain and illustrate why there are two values of θ_A that have only one corresponding value of θ_B. Explain and illustrate why all other values of θ_A have two corresponding values of θ_B.

(b) Show the behavior of the forward kinematics map for Watt's Parallel Motion by picking a variety of points on the configuration space diagram and demonstrating where they map on the operational space figure eight. Among your points, include the two points in the configuration space that map to the crossing point in the figure eight.

Chapter 5

Metric Spaces

One of the most common and useful types of topological space is the so-called metric space. Metric spaces are topological spaces that result from having a means for measuring distance between points in the underlying set. This notion of measuring distance goes beyond stretching out a measuring tape to see how far apart two objects are. For example, as we will see, we can measure the distance between two functions by considering the area bounded between their graphs, and we can measure the distance between two words by considering how many letter changes take us from one to the other. The ability to measure and compare distances between elements of a set is often crucial, and it provides more structure than a general topological space possesses.

Metric spaces play a major role in the mathematical field of analysis, and they appear in a variety of interesting applications. In Section 5.2, we present applications to error-correcting codes and DNA sequences.

First, we begin the chapter with an introduction to metrics, in Section 5.1. After discussing the applications in Section 5.2, we examine properties of metric spaces in Section 5.3 and introduce the concept of metrizability in Section 5.4.

5.1 Metrics

DEFINITION 5.1. *A **metric** on a set X is a function $d : X \times X \rightarrow \mathbb{R}$ with the following properties:*

(i) $d(x, y) \geq 0$ *for all* $x, y \in X$; *equality holds if and only if* $x = y$.

(ii) $d(x, y) = d(y, x)$ *for all* $x, y \in X$.

(iii) $d(x, y) + d(y, z) \geq d(x, z)$ *for all* $x, y, z \in X$ *(the triangle inequality).*

*We call $d(x, y)$ the **distance between x and y**, and we call the pair (X, d), consisting of the set X and the metric d, a **metric space**.*

Note that d has the properties we expect when we measure distance between points. The distance between two points is at least 0, and it equals 0 only when the two points are the same. The distance from point x to point y is the same as the distance from point y to point x. Finally, the distance to travel from x to y and then y to z is never shorter than the distance to travel directly from x to z.

EXAMPLE 5.1. On $\mathbb{R}$, define $d(x, y) = |x - y|$. This is called the **Euclidean metric** or **standard metric** on $\mathbb{R}$. Conditions (i) and (ii) for a metric are immediate. The triangle inequality can be easily verified by considering separately each of the orderings of the three points x, y and z. For instance, in the case that $y \leq x \leq z$, we have

$$
\begin{aligned}
d(x, y) + d(y, z) &= (x - y) + (z - y) \\
&\geq z - y \\
&\geq z - x \\
&= d(x, z).
\end{aligned}
$$

EXAMPLE 5.2. We introduce three different metrics defined on the plane $\mathbb{R}^2$. In Figure 5.1, we illustrate how each measures the distance between two points $p = (p_1, p_2)$ and $q = (q_1, q_2)$.

In Section 0.4 we defined

$$
d(p, q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2}
$$

as the Euclidean distance formula. There we indicated that d satisfies the three properties that make it a metric on the plane. We call d the **Euclidean metric** or **standard metric** on $\mathbb{R}^2$. This metric measures the straight-line distance between points in the plane.

Next, define $d_T(p, q) = |p_1 - q_1| + |p_2 - q_2|$. It is straightforward to show that d_T satisfies the properties of a metric. (See Exercise 5.1.) This metric is called the **taxicab metric** or the **Manhattan metric**, since it measures the total distance traveled vertically together with the total distance traveled horizontally, as if you were restricted to travel on a city grid of parallel and perpendicular streets running North–South and East–West.

Finally, define $d_M(p, q) = \max\{|p_1 - q_1|, |p_2 - q_2|\}$. The function d_M is a metric. (See Exercise 5.2.) It is called the **max metric**. In this case, the distance between two points is the maximum of the differences between their coordinates.
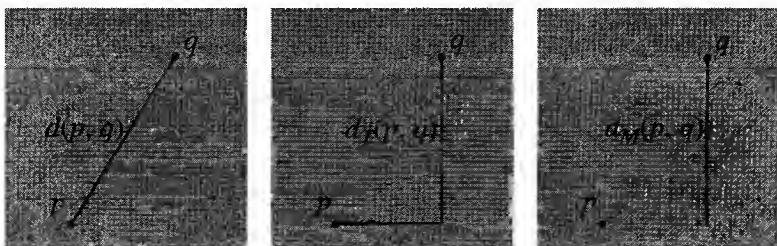
FIGURE 5.1: Measuring distance in the Euclidean, taxicab, and max metrics.

Earlier, when we presented the standard topology on the plane, we defined it via a basis of open balls using the Euclidean distance formula. The proof that the collection of open balls is a basis depended only on the fact that the Euclidean distance formula satisfies the properties of a metric. Therefore, as we show next, given a metric on a set, we can define a topology on the set via open balls that are determined by the metric.

DEFINITION 5.2. *Let (X, d) be a metric space. For $x \in X$ and $\varepsilon > 0$ define the **open ball of radius** ε **centered at** x to be the set*

$$B_d(x, \varepsilon) = \{y \in X \mid d(x, y) < \varepsilon\},$$

*and define the **closed ball of radius** ε **centered at** x to be the set*

$$\bar{B}_d(x, \varepsilon) = \{y \in X \mid d(x, y) \le \varepsilon\}.$$

THEOREM 5.3. *Let (X, d) be a metric space. The collection of open balls, $\mathcal{B} = \{B_d(x, \varepsilon) \mid x \in X, \ \varepsilon > 0\}$, is a basis for a topology on X.*

Before we prove Theorem 5.3, we need the following lemma that shows that, if a point y lies in some open ball $B_d(x, \varepsilon)$, then an open ball centered at y lies in $B_d(x, \varepsilon)$ as well:

LEMMA 5.4. *Let (X, d) be a metric space. If $x \in X$, $\varepsilon > 0$, and $y \in B_d(x, \varepsilon)$, then there exists $\delta > 0$ such that $B_d(y, \delta) \subset B_d(x, \varepsilon)$.*

   Proof. Set δ equal to $\varepsilon - d(x, y)$, as illustrated in Figure 5.2. We claim that $B_d(y, \delta) \subset B_d(x, \varepsilon)$.
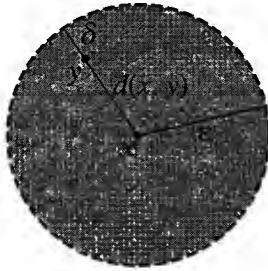
FIGURE 5.2: Set δ equal to $\varepsilon - d(x, y)$.

   To prove the claim, let $z \in B_d(y, \delta)$ be arbitrary; then $d(y, z) < \delta$. Hence,

$$\begin{aligned} d(x, y) + d(y, z) &< d(x, y) + \delta \\ &< d(x, y) + (\varepsilon - d(x, y)) \\ &= \varepsilon. \end{aligned}$$

Thus $d(x, z) < \varepsilon$, implying that $z \in B_d(x, \varepsilon)$. Therefore $B_d(y, \delta) \subset B_d(x, \varepsilon)$. ∎

Proof of Theorem 5.3. We need to check that $\mathcal{B}$ is indeed a basis. It is certainly true that every point x in X is contained in a set in $\mathcal{B}$. In fact, $x \in B_d(x, \varepsilon)$ for every $\varepsilon > 0$.

To see that the second condition for a basis is satisfied, we must show that if $x \in B_1 \cap B_2$, and $B_1, B_2 \in \mathcal{B}$, then there exists $B_3 \in \mathcal{B}$ such that $x \in B_3 \subset B_1 \cap B_2$.

Let B_1 and B_2 be two sets in $\mathcal{B}$, and suppose that $x \in B_1 \cap B_2$. Then by Lemma 5.4 there exist $\delta_1, \delta_2 > 0$ such that $B_d(x, \delta_1) \subset B_1$ and $B_d(x, \delta_2) \subset B_2$. Let $\delta = \min\{\delta_1, \delta_2\}$. Then $x \in B_d(x, \delta) \subset B_1 \cap B_2$, as desired.

It follows that $\mathcal{B}$ is a basis for a topology on X. ∎

Since the collection of open balls associated to a metric is a basis, we immediately obtain a topology:

DEFINITION 5.5. *Let (X, d) be a metric space. The topology generated by the basis of open balls $\mathcal{B} = \{B_d(x, \varepsilon) \mid x \in X, \varepsilon > 0\}$ is called the **topology induced by d** and is referred to as a **metric topology**.*

From now on, when we refer to a metric space (X, d), we assume it is a topological space with the metric topology induced by d.

The following theorem provides a useful condition for a set to be an open set in a metric topology:

THEOREM 5.6. *Let (X, d) be a metric space. A set $U \subset X$ is open in the topology induced by d if and only if for each $y \in U$, there is a $\delta > 0$ such that $B_d(y, \delta) \subset U$. (See Figure 5.3.)*

Proof. See Exercise 5.9. ∎
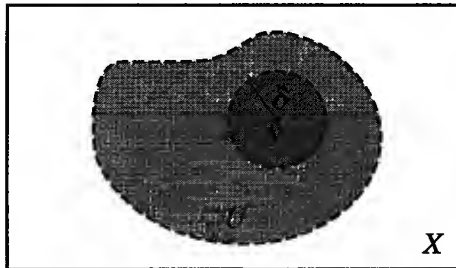
FIGURE 5.3: A set U is open if and only if for each y in U there is an open ball centered at y and contained in U.

In a metric space the open balls are open sets since they are basis elements for the topology. It is straightforward to show that the closed balls are closed sets. (See Exercise 5.14(a).) Furthermore, justifying the notation $\bar{B}_d$ for closed

balls, it follows that the closed ball $\bar{B}_d(x, \varepsilon)$ is the closure of the open ball $B_d(x, \varepsilon)$. (See Exercise 5.14(b).)

EXAMPLE 5.3. Consider the metric on $\mathbb{R}$ given by $d(x, y) = |x - y|$. We see that the basis elements associated with the metric d are the open intervals

$$B_d(x, \varepsilon) = \{y \in \mathbb{R} \mid |x - y| < \varepsilon\} = (x - \varepsilon, x + \varepsilon).$$

Since every open interval (a, b) in the real line can be expressed in the form $(x - \varepsilon, x + \varepsilon)$ by setting $x = \frac{a+b}{2}$ and $\varepsilon = \frac{b-a}{2}$, it follows that this basis is exactly the basis for the standard topology on $\mathbb{R}$. Therefore, the topology induced by the standard metric on $\mathbb{R}$ is the standard topology.

EXAMPLE 5.4. Consider again the three metrics on the plane $\mathbb{R}^2$ introduced in Example 5.2. In Figure 5.4 we illustrate example open balls associated with each of the metrics.
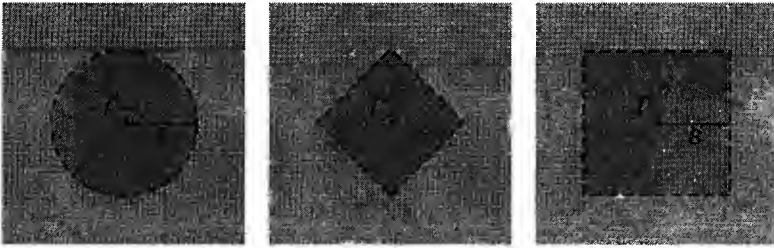
FIGURE 5.4: Open balls in the standard, taxicab, and max metrics.

In the standard metric given by

$$d(p, q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2},$$

the resulting basis elements are the usual round open balls in the plane. In Example 1.11 we defined the topology resulting from this basis as the standard topology on $\mathbb{R}^2$. So the standard topology on $\mathbb{R}^2$ is the topology induced by the standard metric d.

Next, consider the taxicab metric defined by

$$d_T(p, q) = |p_1 - q_1| + |p_2 - q_2|.$$

The resulting basis elements are of the form

$$B_{d_T}(p, \varepsilon) = \{q \in \mathbb{R}^2 \mid |p_1 - q_1| + |p_2 - q_2| < \varepsilon\}.$$

Here, the open ball $B_{d_T}(p, \varepsilon)$ is an open diamond, centered at p, with distance ε from p to the corners. Using Theorem 1.13, it can be shown that this metric also induces the standard topology on $\mathbb{R}^2$. In Section 5.3, we will introduce a general result (Theorem 5.15) that allows for straightforward comparison of

metric topologies; we use it there to prove that the taxicab metric induces the standard topology on $\mathbb{R}^2$.

Finally, consider the max metric defined by

$$d_M(p, q) = \max\{|p_1 - q_1|, |p_2 - q_2|\}.$$

An open ball $B_{d_M}(p, \varepsilon)$ in this metric is an open square in the plane, centered at p, with edge length equal to 2ε. Here too, Theorem 1.13 or Theorem 5.15 can be used to show that the max metric induces the standard topology on $\mathbb{R}^2$. (See Exercise 5.26.)

As with the plane $\mathbb{R}^2$, the Euclidean distance formula on $\mathbb{R}^n$ defines a metric (called the **Euclidean metric** or the **standard metric**) and the resulting metric topology is the standard topology on $\mathbb{R}^n$ that was introduced in Section 1.2.

EXAMPLE 5.5. Let $C[a, b]$ be the set of continuous functions $f : [a, b] \rightarrow \mathbb{R}$. Given two such functions f and g, define

$$\rho(f, g) = \int_a^b |f(x) - g(x)|\, dx.$$

This function measures the area between the graphs of f and g from $x = a$ to $x = b$. (See Figure 5.5.)
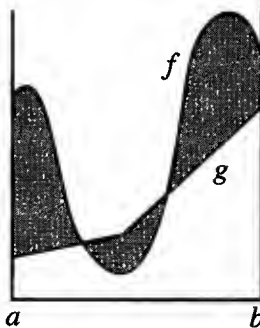
FIGURE 5.5: The distance between f and g is the area between their graphs.

It is straightforward to show that ρ satisfies the properties of a metric. (See Exercise 5.7.)

How do we know $\rho(f, g)$ is defined at all, in particular that the integral in the definition is finite? It turns out that topology provides an answer. Since f and g are continuous functions, so is the integrand $|f - g|$ in the definition of $\rho(f, g)$. In Chapter 7, which introduces the topological property of compactness, we will see that on a compact domain (for example, a closed and bounded interval in $\mathbb{R}$) a continuous real-valued function has a maximum and a minimum value. Therefore, if M is the maximum value for $|f - g|$ over $[a, b]$ then

$$0 \le \int_a^b |f(x) - g(x)|\, dx \le M(b - a),$$

and it follows that $\rho(f, g)$ is finite.

The topological space consisting of $C[a, b]$ with the topology induced by ρ is an example of a **function space**. In the field of mathematical analysis, function spaces are used extensively in the study of mappings, called operators, that send functions to functions.

Exercises for Section 5.1

5.1. Show that the taxicab metric on $\mathbb{R}^2$ satisfies the properties of a metric.

5.2. (a) Show that the max metric on $\mathbb{R}^2$ satisfies the properties of a metric.
   (b) Explain why $d(p, q) = \min\{|p_1 - q_1|, |p_2 - q_2|\}$ does not define a metric on $\mathbb{R}^2$.

5.3. For points $p = (p_1, p_2)$ and $q = (q_1, q_2)$ in $\mathbb{R}^2$ define

$$d_V(p, q) = \begin{cases} 1 & \text{if } p_1 \neq q_1 \text{ or } |p_2 - q_2| \geq 1, \\ |p_2 - q_2| & \text{if } p_1 = q_1 \text{ and } |p_2 - q_2| < 1. \end{cases}$$

   (a) Show that d_V is a metric.
   (b) Describe the open balls in the metric d_V.

5.4. For points p and q in the circle S^1, define $d(p, q)$ to equal the minimum nonnegative angle (in radians) needed to rotate the circle so that the points p and q coincide. Prove that d is a metric on S^1.

5.5. Let X be a nonempty set. Define d on $X \times X$ by

$$d(x, y) = \begin{cases} 0 & \text{if } x = y, \\ 1 & \text{if } x \neq y. \end{cases}$$

   Show that d is a metric, and determine the topology on X induced by d.

5.6. Let d be a metric on a finite set X. Prove that the topology on X induced by d is the discrete topology.

5.7. Consider ρ as defined in Example 5.5.
   (a) Use the basic properties of integrals to prove that ρ is a metric.
   (b) Explain why we cannot generally define $\rho(f, g)$ for functions f and g that are continuous on the open interval (a, b).

5.8. On $C[a, b]$, the set of continuous functions $f : [a, b] \to \mathbb{R}$, define $\rho_M(f, g) = \max_{x \in [a,b]}\{|f(x) - g(x)|\}$. Assuming that such a maximum always exists (something we prove in Chapter 7), prove that ρ_M is a metric.

5.9. **Prove Theorem 5.6:** Let (X, d) be a metric space. A set $U \subset X$ is open in the topology induced by d if and only if for each $y \in U$, there is a $\delta > 0$ such that $B_d(y, \delta) \subset U$.

5.10. (a) Let (X, d) be a metric on a space. For $x, y \in X$, define

$$D(x, y) = \frac{d(x, y)}{1 + d(x, y)}.$$

   Show that D is also a metric on X.
   (b) Explain why no two points in X are distance one or more apart in the metric D.

5.11. On the set of integers $\mathbb{Z}$, show that the function d, defined as follows, is a metric:

$$d(x, y) = \begin{cases} 0 & \text{if } x = y, \\ \min\{ \frac{1}{n!} \mid n! \text{ divides } |x - y| \} & \text{if } x \neq y. \end{cases}$$

5.12. Let S be the set of sequences of 0s and 1s. For $x = (x_1, x_2, x_3, \ldots)$ and $y = (y_1, y_2, y_3, \ldots)$ define

$$d(x, y) = \sum_{j=1}^{\infty} \frac{|x_j - y_j|}{2^j}.$$

 (a) Explain why the infinite sum in the definition of $d(x, y)$ converges for all x and y.
 (b) Prove that $d(x, y)$ is a metric.
 (c) Let E be the subset of S consisting of all sequences that are eventually 0. Thus, $x = (x_1, x_2, x_3, \ldots)$ is in E if there exists $N \geq 0$ such that $x_n = 0$ for all $n \geq N$. Prove that E is dense in S under the topology induced by d.

5.13. Let (X, d) be a metric space. Prove that the distance function $d : X \times X \to \mathbb{R}$ is continuous, assuming that $X \times X$ has the product topology that results from each copy of X having the topology induced by d.

5.14. Let (X, d) be a metric space.
 (a) Show that the closed balls in the metric d are closed sets in the topology on X induced by d.
 (b) Show that for $\varepsilon > 0$ and $x \in X$, the closed ball $\bar{B}_d(x, \varepsilon)$ is the closure of the open ball $B_d(x, \varepsilon)$.

5.15. Let (X, d) be a metric space and assume that $A \subset X$. Prove that $x \in \text{Cl}(A)$ if and only if there exists a sequence in A converging to x.

5.2   Metrics and Information

Metric spaces are used in numerous applications involving the storage, manipulation, and presentation of information. Strings of symbols, like the letters making up the words you read here, are the basic information units. In any situation where we wish to measure the similarities and differences between information units, an appropriate metric can be found to do so. In this section we look at two particular examples. In the first, the information units are transmitted binary codes, and in the second the information units are the sequences of letters modeling nucleotide strands in a DNA molecule.

Error-Correcting Codes

With the incredible amounts of information being transmitted over phone lines, through the internet, or from satellites in space to Earth, it is extremely important to know whether a given message has arrived intact. We expect that there will be some errors in transmission due to electrical surges, cosmic radiation, or a variety of other factors. We want to be able to recognize when this occurs and to correct the faulty message. This brings us to the theory of error-correcting codes.

Suppose that we want to send a certain message. We assume that the message has been encoded in a binary code, which is to say that our message consists of a finite sequence of 0s and 1s, say n of them, which we call a **word**. We also assume that in transmission, some number of 0s may be turned into 1s and vice versa. We do not allow for lost entries or additional entries, so the word that arrives also has length n. Note that what we call a word here could consist of a few words, as we normally think of them, making up a particular message.

Each word of length n can be thought of as a vector of length n, with all entries either 0s or 1s. We write the set of all these possibilities as $V^n = \{(a_1, \ldots, a_n) \mid a_i \in \{0, 1\}\}$. So V^n is the product of n copies of the set $\{0, 1\}$. We now put a metric on this set.

DEFINITION 5.7. *The Hamming distance $D_H(x, y)$ between two words of length n is the number of places in which the words differ.*

For example, given

$$x = (0, 0, 1, 1, 0, 0, 0, 1, 0),$$
$$y = (0, 1, 0, 1, 0, 0, 1, 1, 0),$$

we find that x and y differ in the second, third and seventh places, and therefore $D_H(x, y) = 3$.

The Hamming distance is a metric on V^n. (See Exercise 5.16.) Since V^n is a finite set, the topology that the Hamming distance induces is the discrete topology. (See Exercise 5.6.)

In reality, we do not use all of the elements of V^n to send messages, any more than we use every possible string of n letters as a word in the English language. We take a subset of the words in V^n, and from that we pick the words we transmit.

DEFINITION 5.8. *A **code of length** n is any subset C of V^n. We call the elements of C the **codewords**.*

If the sender and receiver have agreed on a particular code, then when a word arrives that is not one of the codewords, the receiver knows that at least one error has occurred in transmission.

DEFINITION 5.9. *Let C be a code of length n. Define the **minimum distance** of the code C to be the least Hamming distance between two codewords in the code.*

EXAMPLE 5.6. Consider the code of length 6 given by

$C = \{(0, 0, 1, 0, 0, 0), (1, 0, 0, 1, 1, 1), (1, 1, 1, 0, 1, 1), (0, 1, 0, 0, 1, 0)\}$.

You can check that the minimum distance between these codewords is 3.

If we receive a message that we know has at most one error, meaning one change in one of its entries, then we can tell which codeword it is supposed to be. There is no word in V^n that is within one error of two different codewords

in C. If there was, then those two codewords would be within a distance of 2 of one another by the triangle inequality, but that contradicts the fact that the minimum distance between any two codewords is 3.

This demonstrates the basic notion of error-correcting codes. We place an open ball of integer radius r around each codeword. If a word is sent to us and $r - 1$ or fewer errors occur in the transmission, then the word we receive lies in the open ball of radius r centered about the codeword that was originally sent, though it may lie in open balls of radius r centered around other codewords. However, if the open balls of radius r about the different codewords do not overlap, then the word we receive must be in a unique open ball, and we know exactly which codeword was sent. We can correct the errors and determine the intended word. The following theorem makes this rigorous. Let $\lfloor x \rfloor$ denote the largest integer less than or equal to x.

THEOREM 5.10. *If a code C of length n is chosen so that its minimum distance is d, then every message of length n with $\left\lfloor \frac{d-1}{2} \right\rfloor$ or fewer errors can be corrected.*

Proof. Suppose c is the original codeword that was sent and f is the word that arrives. Since f has no more than $\left\lfloor \frac{d-1}{2} \right\rfloor$ errors, we know that $D_H(f, c) \leq \left\lfloor \frac{d-1}{2} \right\rfloor$. Suppose that there is a second codeword c' such that $D_H(f, c') \leq \left\lfloor \frac{d-1}{2} \right\rfloor$. Then we find that

$$
\begin{aligned}
D_H(c, c') &\leq D_H(c, f) + D_H(f, c') \\
&\leq \left\lfloor \frac{d-1}{2} \right\rfloor + \left\lfloor \frac{d-1}{2} \right\rfloor \\
&= 2 \left\lfloor \frac{d-1}{2} \right\rfloor \\
&= d - 1,
\end{aligned}
$$

a contradiction to our minimum distance of d. Hence, c is the only codeword that is within $\left\lfloor \frac{d-1}{2} \right\rfloor$ of f, and therefore must be the codeword that was sent. ∎

There are two useful attributes we would like a code to possess. First, we want it to allow us to correct a large number of errors. This means that we do not want to pick too many codewords, since we would like there to be large open balls that are disjoint from each other, each centered at a codeword. Second, we would like to have enough codewords to enable the sending of a variety of different messages. For example, the code $C = \{(0, 0, 0, 0, 0)\}$ lets us correct up to five errors, but it does not allow the transmission of useful information since it is the only word we can send. Ultimately, we want to maximize the number of disjoint open balls of a given size that we can fit in V^n. (See Exercise 5.19, for example.)

Detecting and correcting errors are active areas of research in the fields of computer science and information science. Metrics, such as the Hamming distance, naturally play a key role in this work.

DNA Sequences

DNA is a long thin molecule made up of millions of atoms. Within its structure lies the code that determines our genetic makeup. Like RNA (introduced in Section 1.4), DNA is composed of nucleotides. While an RNA molecule consists of a single chain of nucleotides, a DNA molecule consists of two chains wound together to form the familiar double helix, as illustrated in Figure 5.6. The nucleotides in DNA come in four types: adenine (A), cytosine (C), guanine (G), and thymine (T). In our introduction to RNA we showed that nucleotides in a chain tend to pair together, contorting the chain into a folded shape. Nucleotides in a DNA chain also pair, but do so with their neighbors on the opposite chain (adenine pairs with thymine, and guanine pairs with cytosine). In fact, the two chains are constructed such that every nucleotide on one chain pairs with its neighbor on the other. Thus the sequence of nucleotides on one chain determines the sequence on the opposite chain, and we can represent part or all of a DNA molecule with a sequence of the letters A, C, G, and T, corresponding to the sequence of nucleotides found in one of the two chains.
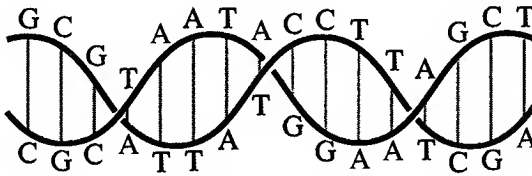
FIGURE 5.6: Diagram of a DNA molecule.

One of the most important problems in DNA research is how to compare distinct DNA sequences. How different is one sequence of DNA from another? In some sense, this is a measure of the evolutionary distance between the two sequences (and, by extension, between the organisms from which they were derived). When a species splits into two new species, resulting in a fork in the evolutionary tree, the species' initially identical DNA sequences begin to accumulate unique changes. Measuring the distance between the two sequences as a function of these differences provides insight into the nature of the evolutionary history of each species.

During the course of evolution, DNA sequence differences arise in a variety of ways. One of the most common is nucleotide substitution, the apparent replacement of a letter in one DNA sequence relative to the original sequence. If this is the only type of change that has occurred between two sequences, then the Hamming distance provides a useful measure of the distance between them (by counting the total number of replacements). Another commonly occurring change in DNA is the insertion or deletion of nucleotides, realized as the insertion or deletion of letters in the corresponding DNA sequence. In this case, all of the subsequent letters in the altered sequence appear offset relative to the original sequence. This results in a large Hamming distance between the two DNA sequences, when in reality they are quite similar. To deal with this problem, we introduce another metric that is useful in making comparisons.

Let x and y be two sequences of the letters A, C, G, and T. We measure the distance between x and y by determining how many operations on x are necessary to turn it into y. We allow three types of operations on x: We can insert any letter into x, we can delete any letter in x, and we can replace any letter with a different letter. For our particular x and y, we can use a sequence S of these operations to turn x into y. Let i_S represent the number of insertions in the sequence, d_S the number of deletions, and r_S the number of replacements. So the total number of operations to turn x into y is $i_S + d_S + r_S$. But of course, there are many different choices of sequences of operations to turn x into y, and therefore we define the distance between x and y as follows:

DEFINITION 5.11. *The Levenshtein distance between sequences x and y is given by*

$$D_L(x, y) = \min_S \{i_S + d_S + r_S\}$$

where the minimum is taken over all sequences S that turn x into y.

EXAMPLE 5.7. Let $x = $ AGTTCGAATCC and $y = $ AGCTCAGGAATC. Then we can get from x to y by the following process:

x	:	AGTTCGAATCC
Replace T	:	AGCTCGAATCC
Insert A	:	AGCTCAGAATCC
Insert G	:	AGCTCAGGAATCC
Delete C	:	AGCTCAGGAATC

We can check, by examining all possibilities with three or fewer operations, that the fewest number of insertions, deletions, and replacements to get us from $x = $ AGTTCGAATCC to $y = $ AGCTCAGGAATC is four, as seen here. Therefore $D_L(x, y) = 4$.

EXAMPLE 5.8. Let $x = $ ACGTTGAATAC and $y = $ AGGGTTGAATA. Visually inspecting x and y, we see that they appear somewhat similar. In fact, they have the segments GTTGAATA in common. It is not difficult to determine that the Levenshtein distance between x and y is 3. In contrast, if we compute the Hamming distance between x and y (counting the number of entries where they differ), we obtain 7. Thus, in this example we see that in comparison with the Hamming distance, the Levenshtein distance better reflects the proximity of x and y resulting from their similar structure.

The Levenshtein distance is also used for spell checking, speech recognition, and plagiarism detection. It is just one of a variety of metrics researchers are using for DNA sequence comparison and analysis.

Exercises for Section 5.2

5.16. Show that the Hamming distance is a metric on V^n, the set of all words of length n.

5.17. Can you extend the Hamming distance to be a metric on words of any length? That is to say, is there a way to extend the definition to obtain a metric that allows a distance between two words of differing length but that yields the same distance as the Hamming distance when applied to two words of the same length?

5.18. Compare the Hamming metric and the Levenshtein metric on the set V^n. Is either metric, in general, always less than or equal to the other? Prove your assertion.

5.19. We say that a code of length p **corrects** n **errors** if every word in V^p is within distance n of a unique codeword in the code.
   (a) In V^8, find an example of a code C that has four codewords and corrects two errors.
   (b) Show that four codewords is the maximal size for a code in V^8 that corrects two errors.

5.20. Let W be the set of sequences of finite length made up of the letters A, C, G, and T. Show that the Levenshtein distance is a metric on W.

5.21. For each of the following, find the Levenshtein distance between the two sequences, x and y. In cases where the sequences have equal length, also compute the Hamming distance.
   (a) $x =$ ACGGTAT and $y =$ GGTAG
   (b) $x =$ CTGGTAC and $y =$ CTAGATC
   (c) $x =$ CCAGTCA and $y =$ CCGTCTTA
   (d) $x =$ TGACCGTTA and $y =$ TGCGCTTAG

5.22. A word was badly misspelled as TUPOTAGRY. Suppose that a spell checker changes the word to a word in its database that is the least Levenshtein distance away from the misspelling. Which of the following would the spell checker pick as the intended word: TOPOGRAPHY, TOPOLOGY, or TAUTOLOGY?

5.3   Properties of Metric Spaces

Metric spaces have many useful properties, one of which is presented in the following theorem:

THEOREM 5.12. *Every metric space is Hausdorff.*

   Proof. Let (X, d) be a metric space. Suppose x and y are distinct points in X with $d(x, y) = \varepsilon$. Consider the sets $U = B_d(x, \varepsilon/2)$ and $V = B_d(y, \varepsilon/2)$. It follows that $x \in U$, $y \in V$, and U and V are open sets. We claim that U and V are disjoint. Suppose $U \cap V \neq \varnothing$, and z is in the intersection. Then $d(x, z) < \varepsilon/2$ and $d(y, z) < \varepsilon/2$. Therefore, by the triangle inequality,

$$d(x, y) \leq d(x, z) + d(z, y) < \varepsilon/2 + \varepsilon/2 = \varepsilon;$$

   that is, $d(x, y) < \varepsilon$. This contradicts $d(x, y) = \varepsilon$. Thus $U \cap V = \varnothing$. Hence, there exist disjoint open sets U and V containing x and y, respectively, implying that X is Hausdorff. ∎

Theorem 5.12 implies that if a topological space is not Hausdorff, then it cannot be induced by a metric. For example, the topological space $\mathbb{R}_{fc}$, the real numbers with the finite complement topology, is not Hausdorff and therefore cannot be induced by a metric on $\mathbb{R}$. Also, the digital line is not Hausdorff and therefore cannot be induced by a metric on $\mathbb{Z}$. We discuss this idea of a topological space being induced by a metric (being "metrizable") further in Section 5.4.

The $\varepsilon - \delta$ definition of continuity for functions $f : \mathbb{R} \rightarrow \mathbb{R}$ generalizes to functions $f : X \rightarrow Y$ if X and Y are metric spaces. The next theorem establishes that the resulting $\varepsilon - \delta$ definition is equivalent to the topological open set definition.

THEOREM 5.13. *Let* (X, d_X) *and* (Y, d_Y) *be metric spaces. A function* $f : X \rightarrow Y$ *is continuous in the open set definition if and only if for each* $x \in X$ *and* $\varepsilon > 0$, *there exists a* $\delta > 0$ *such that if* $x' \in X$ *and* $d_X(x, x') < \delta$, *then* $d_Y(f(x), f(x')) < \varepsilon$.

Proof. See Exercise 5.24.   ∎

In a metric space, not only can we measure distance between points, but we can also measure distance between sets, as follows:

DEFINITION 5.14. *Let* (X, d) *be a metric space. For sets* $A, B \subset X$ *define the distance between* A *and* B *by*

$$d(A, B) = glb\{d(a, b) \mid a \in A, \ b \in B\}.$$

The greatest lower bound in Definition 5.14 exists for every pair of sets A and B since the set of values $\{d(a, b) \mid a \in A, \ b \in B\}$ is bounded below by 0.

EXAMPLE 5.9. Unlike the situation with points, if the distance between two sets is 0, the sets need not be equal. For example, in the standard metric on $\mathbb{R}^2$, if we let A be the x-axis and B be the y-axis, then $d(A, B) = 0$ but $A \neq B$. In fact, $d(A, B) = 0$ does not necessarily imply that A and B have a point in common. In $\mathbb{R}^2$ again, with the standard metric, if we let A be the x-axis and B be the open ball of radius 1 centered at the point $(0, 1)$, then $d(A, B) = 0$ but $A \cap B = \varnothing$. (See Figure 5.7.)
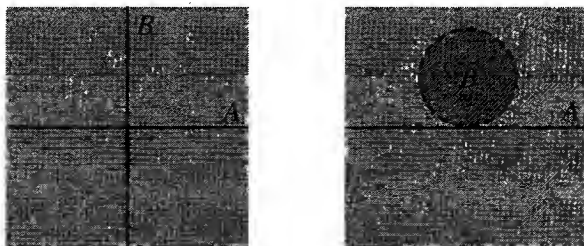
FIGURE 5.7: In both cases $d(A, B) = 0$.

We have already seen that a given set can have many different metrics. It will be useful to be able to compare the topologies they induce. The following theorem provides a means for doing so:

THEOREM 5.15. *Let d and d' be metrics on a set X, and let $\mathcal{T}$ and $\mathcal{T}'$, respectively, be the topologies that they induce. Then $\mathcal{T}'$ is finer than $\mathcal{T}$ if and only if for each $x \in X$ and $\varepsilon > 0$, there exists a $\delta > 0$ such that $B_{d'}(x, \delta) \subset B_d(x, \varepsilon)$. (See Figure 5.8.)*
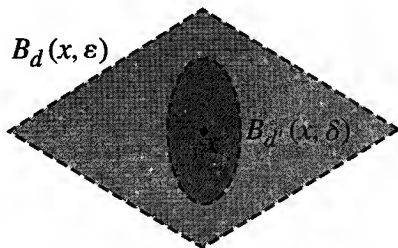
FIGURE 5.8: For each $x \in X$ and $\varepsilon > 0$, there exists δ such that $B_{d'}(x, \delta) \subset B_d(x, \varepsilon)$.

Proof. Suppose that $\mathcal{T}'$ is finer than $\mathcal{T}$. Then every open set in $\mathcal{T}$ is open in $\mathcal{T}'$. In particular, for every $x \in X$ and $\varepsilon > 0$, $B_d(x, \varepsilon)$ is open in $\mathcal{T}$ and hence is open in $\mathcal{T}'$. Since $B_d(x, \varepsilon)$ is open in $\mathcal{T}'$ and contains x, Theorem 5.6 implies that there is a $\delta > 0$ such that $B_{d'}(x, \delta) \subset B_d(x, \varepsilon)$, as we wished to show.

Suppose now that for each $x \in X$ and $\varepsilon > 0$, there exists a $\delta > 0$ such that $B_{d'}(x, \delta) \subset B_d(x, \varepsilon)$. We prove that $\mathcal{T}'$ is finer than $\mathcal{T}$. Let U be an open set in $\mathcal{T}$. We show that U is open in $\mathcal{T}'$. Let x be an arbitrary point in U. Since U is open in $\mathcal{T}$, Theorem 5.6 implies that there is an $\varepsilon > 0$ such that $B_d(x, \varepsilon) \subset U$. By assumption there is a $\delta > 0$ such that $B_{d'}(x, \delta) \subset B_d(x, \varepsilon) \subset U$. It follows that for each $x \in U$ there exists $\delta > 0$ such that $B_{d'}(x, \delta) \subset U$. Theorem 5.6 implies that U is open in $\mathcal{T}'$, as we wished to show. ∎

Using Theorem 5.15, we can easily prove that the standard metric, the taxicab metric, and the max metric all induce the same topology on $\mathbb{R}^2$. In the next theorem we prove that the first two of these metrics induce the same topology on $\mathbb{R}^2$. In Exercise 5.26 we ask you to show that the topologies induced by the taxicab metric and the max metric are the same.

As we indicated in Section 1.2, the standard topology on the plane is not dependent on the particular shape of the basis elements; the round, open balls of the standard metric, the open diamonds of the taxicab metric, and the open squares of the max metric all generate the standard topology on the plane.

THEOREM 5.16. *The standard metric and the taxicab metric induce the same topology on $\mathbb{R}^2$.*

Proof. We begin by showing that the topology induced by the standard metric d is finer than the topology induced by the taxicab metric d_T. Let $p \in \mathbb{R}^2$ and $\varepsilon > 0$ be arbitrary. Set δ equal to $\varepsilon/2$. We claim that $B_d(p, \delta) \subset B_{d_T}(p, \varepsilon)$. Given the claim, it then follows by Theorem 5.15 that the topology induced by d is finer than the topology induced by d_T.

To prove the claim, let $q \in B_d(p, \delta)$ be arbitrary. Note that

$$|p_1 - q_1| = \sqrt{(p_1 - q_1)^2} \leq \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2} = d(p, q).$$

Since $d(p, q) < \delta = \varepsilon/2$, it follows that $|p_1 - q_1| < \varepsilon/2$. Similarly, $|p_2 - q_2| < \varepsilon/2$. Thus, $d_T(p, q) = |p_1 - q_1| + |p_2 - q_2| < \varepsilon$, and therefore $q \in B_{d_T}(p, \varepsilon)$. Hence, $B_d(p, \delta) \subset B_{d_T}(p, \varepsilon)$, as we wished to show.

Next, we show that the topology induced by d_T is finer than the topology induced by d. Let $p \in \mathbb{R}^2$ and $\varepsilon > 0$ be arbitrary. Here we set $\delta = \varepsilon/\sqrt{2}$. We claim that $B_{d_T}(p, \delta) \subset B_d(p, \varepsilon)$, implying the desired result by Theorem 5.15.

Thus let $q \in B_{d_T}(p, \delta)$ be arbitrary. Note that

$$|p_1 - q_1| \leq |p_1 - q_1| + |p_2 - q_2| < \delta.$$

Similarly, $|p_2 - q_2| < \delta$. Therefore,

$$d(p, q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2} < \sqrt{\delta^2 + \delta^2} = \delta\sqrt{2} = \varepsilon.$$

Hence, $q \in B_d(p, \varepsilon)$, implying that $B_{d_T}(p, \delta) \subset B_d(p, \varepsilon)$ as claimed. ∎

Of course, not all metrics on the plane induce the standard topology on the plane. The function d_V defined by

$$d_V(p, q) = \begin{cases} 1 & \text{if } p_1 \neq q_1 \text{ or } |p_2 - q_2| \geq 1, \\ |p_2 - q_2| & \text{if } p_1 = q_1 \text{ and } |p_2 - q_2| < 1, \end{cases}$$

is a metric that induces a topology on the plane that is strictly finer than the standard topology. (See Exercises 5.3 and 5.27.)

We next show how we can dramatically alter a metric and still leave the topology it induces unchanged.

DEFINITION 5.17. *Let (X, d) be a metric space. A subset A of X is said to be **bounded under** d if there exists a $\mu > 0$ such that $d(x, y) \leq \mu$ for all $x, y \in A$. If X itself is bounded under d, then we say that d is a **bounded metric**.*

Notice that with a bounded metric on a set X, every subset of X is itself bounded.

Amazingly enough, whether or not a metric is bounded does not have implications for the topology that it induces. The next theorem shows that every topology induced by a metric is induced by a bounded metric.

THEOREM 5.18. *Let (X, d) be a metric space, and define $d' : X \times X \to \mathbb{R}$ by $d'(x, y) = min\{d(x, y), 1\}$. Then d' is a bounded metric that induces the same topology as d.*

> **Proof.** Since d' cannot have a value greater than 1, it immediately follows that d' is bounded. But we do need to check that d' is a metric. Clearly, $d'(x, y) \geq 0$ for all x and y in X. We also see that $d'(x, y) = 0$ if and only if $d(x, y) = 0$, which happens exactly when $x = y$. In addition, $d'(x, y) = d'(y, x)$, since the same holds true for d. Hence, all we have left to show is the triangle inequality.
>
> To prove the triangle inequality, we consider two cases for $x, y, z \in X$. First, suppose that $d(x, y) \geq 1$ or $d(y, z) \geq 1$. Then, it follows that $d'(x, y) + d'(y, z) \geq 1$. But $d'(x, z) \leq 1$, so we have $d'(x, y) + d'(y, z) \geq d'(x, z)$. On the other hand, suppose $d(x, y) < 1$ and $d(y, z) < 1$. Then,
>
> $$d'(x, y) + d'(y, z) = d(x, y) + d(y, z) \geq d(x, z) \geq d'(x, z).$$
>
> Thus, d' satisfies the triangle inequality.
>
> We use Theorem 5.15 to show that the topologies $\mathcal{T}$ and $\mathcal{T}'$, induced by d and d', respectively, are the same. First, we show that $\mathcal{T}$ is finer that $\mathcal{T}'$. Let $x \in X$ and $\varepsilon > 0$ be arbitrary. We consider two cases for ε. First, if $\varepsilon \leq 1$, then $B_d(x, \varepsilon) = B_{d'}(x, \varepsilon)$. However, if $\varepsilon > 1$, then $B_{d'}(x, \varepsilon) = X$, and therefore $B_d(x, \varepsilon) \subset B_{d'}(x, \varepsilon)$ clearly holds. In either case, if we let $\delta = \varepsilon$, then we have $B_d(x, \delta) \subset B_{d'}(x, \varepsilon)$. It follows that $\mathcal{T}$ is finer than $\mathcal{T}'$.
>
> Next we show that $\mathcal{T}'$ is finer than $\mathcal{T}$. Again take an arbitrary $x \in X$ and $\varepsilon > 0$. As above, if $\varepsilon \leq 1$, then $B_d(x, \varepsilon) = B_{d'}(x, \varepsilon)$. If $\varepsilon > 1$, then $B_{d'}(x, 1) \subset B_d(x, \varepsilon)$. Thus, if we set $\delta = min\{\varepsilon, 1\}$, then $B_{d'}(x, \delta) \subset B_d(x, \varepsilon)$ holds for all $x \in X$ and $\varepsilon > 0$. Therefore $\mathcal{T}'$ is finer than $\mathcal{T}$.
>
> It now follows that the topologies induced by d and d' are the same. ∎

EXAMPLE 5.10. If we replace the standard metric $d(x, y) = |x - y|$ on $\mathbb{R}$ by the bounded metric $d'(x, y) = min\{1, |x - y|\}$, we again obtain the standard topology. But now our basis elements are open intervals of length at most 2 and the set $\mathbb{R}$ itself. We obtain $\mathbb{R}$ as a basis element because $B_{d'}(x, \gamma) = \mathbb{R}$ for all $\gamma > 1$ and $x \in \mathbb{R}$.

In Theorem 5.18 there is nothing significant about the value 1 as the cutoff for defining a bounded metric that induces the same topology as the metric d. We could cut d off at one-millionth and still obtain the same topology. In fact, Theorem 5.18 holds if we replace the bound of 1 with a bound given by an arbitrary $\varepsilon > 0$. In some sense the topology induced by the metric only depends on what is really close together in the metric.

Next, we define a notion of equivalence for metrics. The idea is that two metric spaces are equivalent if there is a bijection between them that preserves distance. Specifically, we have the following definition:

DEFINITION 5.19. *Let* (X, d_X) *and* (Y, d_Y) *be metric spaces. A bijective function* $f : X \to Y$ *is called an* **isometry** *if* $d_X(x, x') = d_Y(f(x), f(x'))$ *for every pair of points* x *and* x' *in* X. *If* $f : X \to Y$ *is an isometry, then we say that the metric spaces* X *and* Y *are* **isometric.**

In this definition it suffices to use a surjective function f in place of the bijective function f. If a function f preserves distance, then f is injective. (See Exercise 5.30.) Therefore f being surjective and distance preserving implies that f is bijective. We use bijective in the definition to emphasize the fact that an isometry must be both bijective and distance preserving.

Just as homeomorphism is the fundamental equivalence between topological spaces, isometry is the fundamental equivalence between metric spaces. All of the properties of a given metric space are preserved by isometry. Two metric spaces that are isometric are indistinguishable except for the particular names on the elements.

Isometry is a stronger form of equivalence than homeomorphism. In particular, isometric spaces are homeomorphic (see Exercise 5.31) but not vice versa, as the following example indicates:

EXAMPLE 5.11. The plane with the taxicab metric d_T is not isometric to the plane with the standard metric d, even though these two metrics induce the same topology. To see this, suppose that we have an isometry $f : \mathbb{R}^2 \to \mathbb{R}^2$ such that $d(f(p), f(q)) = d_T(p, q)$ for all $p, q \in \mathbb{R}^2$. Consider the four points $\alpha = (0, 0)$, $\beta = (1, 0)$, $\gamma = (0, 1)$ and $\delta = (1, 1)$. In the taxicab metric, $d_T(\alpha, \beta) = 1$, $d_T(\alpha, \gamma) = 1$, $d_T(\beta, \delta) = 1$, $d_T(\gamma, \delta) = 1$, $d_T(\alpha, \delta) = 2$, and $d_T(\beta, \gamma) = 2$, as illustrated in Figure 5.9. Since f is an isometry, in the standard metric we must have $d(f(\alpha), f(\beta)) = 1$, $d(f(\alpha), f(\gamma)) = 1$, and $d(f(\beta), f(\gamma)) = 2$. However, in the standard metric this can only occur if $f(\beta)$, $f(\gamma)$, and $f(\alpha)$ lie in a line with $f(\alpha)$ equidistant from $f(\beta)$ and $f(\gamma)$. Similarly, $f(\beta)$, $f(\gamma)$, and $f(\delta)$ must lie in a line with $f(\delta)$ equidistant from $f(\beta)$ and $f(\gamma)$. Hence, $f(\alpha) = f(\delta)$. But this contradicts f being a bijection, and therefore no such isometry exists.
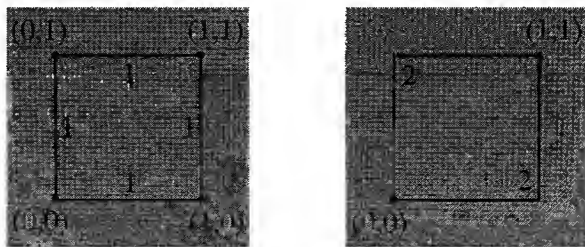
FIGURE 5.9: Distances between points in the taxicab metric.

Exercises for Section 5.3

5.23. Let (X, d) be a metric space. Let A and B be disjoint subsets of X that are closed in the topology induced by d. Prove that there exist disjoint open sets U and V such that $A \subset U$ and $B \subset V$.

5.24. Prove Theorem 5.13: Let (X, d_X) and (Y, d_Y) be metric spaces. A function $f : X \to Y$ is continuous in the open set definition if and only if for each $x \in X$ and $\varepsilon > 0$, there exists a $\delta > 0$ such that if $x' \in X$ and $d_X(x, x') < \delta$, then $d_Y(f(x), f(x')) < \varepsilon$. (Hint: Consider Exercise 4.3 and the proof of Theorem 4.6.)

5.25. Let (X, d) be a metric space, and assume $p \in X$ and $A \subset X$.
   (a) Provide an example showing that $d(\{p\}, A) = 0$ need not imply that $p \in A$.
   (b) Prove that if A is closed and $d(\{p\}, A) = 0$, then $p \in A$.

5.26. Use Theorem 5.15 to prove that the taxicab metric and the max metric induce the same topology on $\mathbb{R}^2$.

5.27. Consider the metric d_V on $\mathbb{R}^2$ defined by

$$d_V(p, q) = \begin{cases} 1 & \text{if } p_1 \neq q_1 \text{ or } |p_2 - q_2| \geq 1, \\ |p_2 - q_2| & \text{if } p_1 = q_1 \text{ and } |p_2 - q_2| < 1. \end{cases}$$

   (a) Use Theorem 5.15 to prove that the topology induced by d_V is finer than the standard topology.
   (b) Show that the standard topology is not finer than the topology induced by d_V.

5.28. Let (X, d) be a metric space. The function

$$D(x, y) = \frac{d(x, y)}{1 + d(x, y)}$$

is a bounded metric on X. (See Exercise 5.10.) Show that the topologies induced by D and d are the same.

5.29. On the set of continuous functions $C[a, b]$ consider the metrics ρ_M and ρ defined by

$$\rho_M(f, g) = \max_{x \in [a,b]}\{|f(x) - g(x)|\}, \text{ and}$$

$$\rho(f, g) = \int_a^b |f(x) - g(x)| \, dx.$$

These metrics were introduced in Exercise 5.8 and Example 5.5, respectively.
   (a) Use Theorem 5.15 to prove that the topology induced by ρ_M on $C[a, b]$ is finer than the topology induced by ρ.
   (b) Show that for every $c_1, c_2 > 0$ there exists $f \in C[a, b]$ such that $\max_{x \in [a,b]}\{|f(x)|\} = c_1$ and

$$\int_a^b |f(x)| \, dx = c_2.$$

   (c) Let $Z \in C[a, b]$ be the function defined by $Z(x) = 0$ for all $x \in [a, b]$.
         Given $\varepsilon > 0$, show that no $\delta > 0$ exists such that $B_\rho(Z, \delta) \subset B_{\rho_M}(Z, \varepsilon)$.
         (Hint: Part (b) helps.)

   (d) What does Theorem 5.15 allow us to conclude from (c)?

5.30. Let (X, d_X) and (Y, d_Y) be metric spaces. Show that if $f : X \to Y$ is such that
$d_X(x, x') = d_Y(f(x), f(x'))$ for all $x, x' \in X$, then f is injective.

5.31. Let (X, d_X) and (Y, d_Y) be metric spaces and $f : X \to Y$ be an isometry
between them. Show that f is a homeomorphism between the corresponding
metric spaces.

5.32. Show that the max metric on $\mathbb{R}^2$, given by

$$d_M(p, q) = \max\{|p_1 - q_1|, |p_2 - q_2|\},$$

is not isometric to the standard metric on $\mathbb{R}^2$.

5.4 Metrizability

There are many topological spaces that are not induced by a metric. For
instance, Theorem 5.12 implies that a topology that is not Hausdorff cannot
be induced by a metric. But when a topology is induced by a metric, we
have some immediate additional structure that can be very useful. Thus, given
a topological space, we are interested in knowing if it can be induced by a
metric.

DEFINITION 5.20. *Let X be a topological space. We say X is **metrizable** if
there exists a metric d on X that induces the topology on X.*

EXAMPLE 5.12. In Example 3.4 we introduced the standard topology on the
circle S^1 as the subspace topology inherited from the standard topology on the
plane $\mathbb{R}^2$. Is this topology on S^1 metrizable? Yes! First, note that we obtain a
basis for the subspace topology on S^1 by intersecting open balls in the plane
with S^1. The resulting sets are open intervals along the circle.

   Consider the metric on S^1 defined by setting $d(p, q)$ equal to the minimum
nonnegative angle (in radians) needed to rotate the circle so that points p and q
coincide. (See Exercise 5.4.) In this metric, the resulting open balls are open
intervals on the circle. Thus the basis of open balls for the topology on S^1
induced by d is the same as the basis just described for the standard topology
on S^1. It follows that the standard topology on S^1 is metrizable.

   In the previous example we showed that the subspace S^1 of the metric
space $\mathbb{R}^2$ is itself metrizable. This idea holds in general. If X is a metric space
and Y is a subset of X, then the subspace topology on Y is metrizable. (See
Exercise 5.34.)

   Given a set X, the discrete topology on X is metrizable. (See Exer-
cise 5.33.) In fact, if X is finite, then every metric on X induces the discrete

topology. (See Exercise 5.6.) Thus, on a finite set the discrete topology is the only metrizable topology.

Metrizability is a topological property, as the following theorem indicates:

THEOREM 5.21. *If X is a metrizable topological space and Y is homeomorphic to X, then Y is metrizable.*

Proof. See Exercise 5.35. ∎

It turns out that many topological spaces are metrizable, and mathematicians have long been interested in the question, "Can we determine if a topological space is metrizable without explicitly finding a metric for it; that is, are there conditions on the space that ensure that it is metrizable?" The answer is yes. The famous Urysohn Metrization Theorem asserts that if two simple conditions are satisfied (the space is "regular" and has a countable basis), we are guaranteed that the space is metrizable. We discuss this result in the remainder of this section.

The topological property of regularity is a strengthening of the Hausdorff property. Recall that a topological space X is Hausdorff if for every pair of distinct points in X, there exists a pair of disjoint open sets, each containing one of the points.

DEFINITION 5.22. *Let X be a topological space. We say that X is **regular** if*

(i) *One-point sets are closed in X;*

(ii) *For every $a \in X$ and every closed set B in X that does not contain a, there exist disjoint open sets U and V such that $a \in U$ and $B \subset V$. (See Figure 5.10.)*
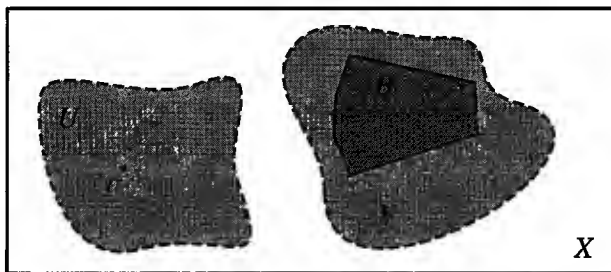
FIGURE 5.10: Separating a point and a closed set in a regular space.

EXAMPLE 5.13. The real line $\mathbb{R}$ with the standard topology is regular. One-point sets are closed in $\mathbb{R}$. Choose a point $x \in \mathbb{R}$ and a closed set C that does not contain x. The complement of C is an open set containing x, and therefore it must contain an open interval (a, b) that contains x. We have $a < x < b$, and we can choose c and d so that $a < c < x < d < b$. Then $U = (-\infty, c) \cup (d, \infty)$ is an open set containing C, and $V = (c, d)$ is an open set containing x. Furthermore, U and V are disjoint, as required.

By combining conditions (i) and (ii) in the definition of regularity, it is straightforward to prove that if a topological space is regular, then it is Hausdorff. However, without the requirement that one-point sets be closed, this need not be the case; specifically, it is possible to have a topological space that satisfies only condition (ii) in Definition 5.22 and that is not Hausdorff. (See Exercise 5.42.)

While being regular implies being Hausdorff, the converse need not hold. In the next example we introduce a topological space that is Hausdorff but not regular. Thus regularity is a stronger property.

EXAMPLE 5.14. On $\mathbb{R}$, consider the collection consisting of all open intervals (a, b) and all subsets $(c, d) \cap \mathbb{Q}$. This is a basis for a topology on $\mathbb{R}$, and the resulting topological space is Hausdorff but not regular. (See Exercise 5.37.)

The properties of being Hausdorff and being regular are known as **separation axioms**. There is another separation axiom worth mentioning, although it does not play a role in the Urysohn Metrization Theorem.

DEFINITION 5.23. *Let X be a topological space. We say that X is **normal** if*

    (i) One-point sets are closed in X;

    (ii) For every pair of disjoint closed sets A and B in X, there exist disjoint open sets U and V such that $A \subset U$ and $B \subset V$. (See Figure 5.11.)
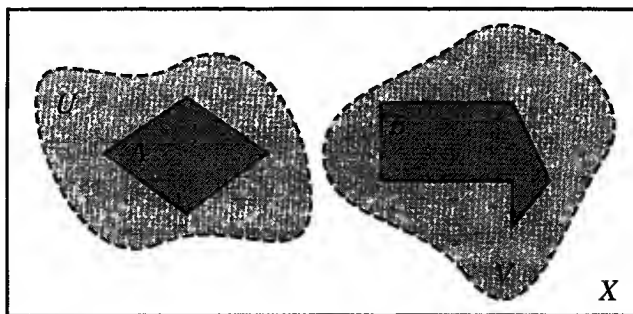
FIGURE 5.11: Separating two closed sets in a normal space.

Not only is $\mathbb{R}$ regular in the standard topology, but it is normal as well. In fact, every metric space is normal. (See Exercise 5.23.) Furthermore, it is an easy consequence of the definitions that every normal space is regular.

The second assumption in the Urysohn Metrization Theorem is that the topological space has a countable basis. Recall that a set is countable if it is either finite or can be put in a one-to-one correspondence with the positive integers.

EXAMPLE 5.15. The real line $\mathbb{R}$ with the standard topology has a countable basis. Consider the collection $C = \{(a, b) \mid a, b \in \mathbb{Q}, a < b\}$. It is a countable collection of intervals since it is indexed by $\mathbb{Q} \times \mathbb{Q}$, a countable set. (See Theorem 0.29.) Using Theorem 1.13, it is straightforward to show that C is a basis for the standard topology on $\mathbb{R}$.

Similarly, the standard topology on $\mathbb{R}^n$ has a countable basis given by the collection of all products of open intervals having rational endpoints.

Pavel Urysohn (1898–1924) was one of the most promising mathematicians of his era. Unfortunately, he drowned at the age of 26 while swimming in rough seas off the coast of France. It was in 1924 that he proved the following theorem that now bears his name:

THEOREM 5.24. *The Urysohn Metrization Theorem. If a topological space X is regular and has a countable basis, then X is metrizable.*

We do not prove the Urysohn Metrization Theorem since its proof requires tools that we have not developed. A proof of the Urysohn Metrization Theorem can be found in [Mun]. The idea behind the proof, however, is straightforward. Using the assumptions that X is regular and has a countable basis, it can be shown that X can be embedded in a metric space. Therefore X is homeomorphic to a subspace of a metric space. Since a subspace of a metric space is metrizable, and since metrizability is a topological property, it follows that X is metrizable.

EXAMPLE 5.16. The arithmetic progression topology on $\mathbb{Z}$ is defined via the basis consisting of all arithmetic progressions

$$A_{a,b} = \{..., a - 2b, a - b, a, a + b, a + 2b, ...\},$$

for $a, b \in \mathbb{Z}$ with $b \neq 0$. (See Exercise 1.15.) This basis is countable since it is a collection indexed by a subset of $\mathbb{Z} \times \mathbb{Z}$, a countable set.

Furthermore, the arithmetic progression topology on $\mathbb{Z}$ is regular. (See Exercise 5.39.)

Since this is a regular topological space with a countable basis, it is metrizable. There must be a metric that induces this topology. In fact, we can find it. The metric on $\mathbb{Z}$ given by

$$d(x, y) = \begin{cases} 0 & \text{if } x = y, \\ \min\{ \frac{1}{n!} \mid n! \text{ divides } |x - y| \} & \text{if } x \neq y, \end{cases}$$

induces the arithmetic progression topology on $\mathbb{Z}$. (See Exercises 5.11 and 5.40.)

The Urysohn Metrization Theorem indicates that every regular topological space with a countable basis is metrizable. Does the converse hold? In other words, must a metrizable space have a countable basis and be regular? The

answer to half of this question is yes. As we indicated above, a metric space is normal, and a normal space is regular; therefore a metric space is regular. However, the following example demonstrates that the answer to the other half of the question is no. A metric space does not necessarily have a countable basis.

EXAMPLE 5.17. The real line $\mathbb{R}$ with the discrete topology is metrizable. (See Exercise 5.33.) Since this topology is discrete, every one-point set is an open set. There are uncountably many such sets since $\mathbb{R}$ is uncountable.

Now, given a basis for a topology, every open set in the topology must be a union of basis elements. It follows that every basis for the discrete topology must include each one-point set as a basis element. Therefore every basis for the discrete topology on $\mathbb{R}$ must contain the uncountable collection of one-point sets. Thus, although $\mathbb{R}$ with the discrete topology is a metrizable space, it does not have a countable basis.

Exercises for Section 5.4

5.33. Let X be a set.
   (a) Show that the discrete topology on X is induced by the metric

$$d(x, y) = \begin{cases} 0 & \text{if } x = y, \\ 1 & \text{if } x \neq y. \end{cases}$$

   (b) Is the trivial topology on X metrizable?

5.34. Let (X, d) be a metric space, and let Y be a subset of X. Consider the function $d_Y : Y \times Y \to \mathbb{R}$, given by $d_Y(x, y) = d(x, y)$.
   (a) Prove that d_Y is a metric on Y.
   (b) Prove that the topology induced by d_Y on Y is the subspace topology that Y inherits from the metric topology on X.

5.35. Let (X, d) be a metric space. Assume that $f : X \to Y$ is a homeomorphism, and define $d^* : Y \times Y \to \mathbb{R}$ by $d^*(x, y) = d(f^{-1}(x), f^{-1}(y))$. (Theorem 5.21 follows from exercises (a) and (b).)
   (a) Prove that d^* is a metric on Y.
   (b) Prove that the topology on Y is induced by the metric d^*.
   (c) Prove that f is an isometry between (X, d) and (Y, d^*).

5.36. Prove that $\mathbb{R}$ with the lower limit topology is a regular topological space.

5.37. On $\mathbb{R}$, consider the collection $\mathcal{B}$ consisting of all open intervals (a, b) and all subsets $(c, d) \cap \mathbb{Q}$.
   (a) Show that $\mathcal{B}$ is a basis for a topology on $\mathbb{R}$.
   (b) Show that $\mathbb{R}$ is Hausdorff in this topology.
   (c) Show that $\mathbb{R}$ is not regular in this topology.

5.38. Show that if X and Y are regular, then so is the product space $X \times Y$. Conclude that $\mathbb{R}^n$ is regular.

5.39. Show that the arithmetic progression topology on $\mathbb{Z}$ is regular.

5.40. Show that the arithmetic progression topology on $\mathbb{Z}$ is induced by the metric

$$d(x, y) = \begin{cases} 0 & \text{if } x = y, \\ \min\{ \frac{1}{n!} \mid n! \text{ divides } |x - y| \} & \text{if } x \neq y. \end{cases}$$

5.41. Show that regularity and normality are topological properties. That is, show that if X is regular and Y is homeomorphic to X, then Y is regular (and do the same for normality).

5.42. Show that the two-point space $X = \{x_1, x_2\}$ with the trivial topology satisfies the second condition in the definition of regularity; that is, show that for every point a in X and every closed set B in X not containing a, there exist disjoint open sets U and V such that $a \in U$ and $B \subset V$. (Note: Since one-point sets are not closed in this topological space, this is not a regular space. Nor is it Hausdorff. This demonstrates why in the definition of a regular topological space the second condition alone does not suffice if we wish to have a regular space be Hausdorff.)

Chapter 6

Connectedness

There are several natural approaches that we could take to rigorously capture the concept of connectedness for a topological space. One approach might be to say that a topological space is connected if it cannot be broken down into two distinct pieces that are separated from each other. Another approach might be to say that a topological space is connected if we can take a continuous walk in the space from any point to any other point. In this chapter we define both of these types of connectedness. Spaces of the first type are called connected, while spaces of the second type are called path connected. We prove that these two types of connectedness are not the same, but that path connectedness does imply connectedness.

As simple as the concept of connectedness appears, it has profound implications for topology and its applications. We show that connectedness is the key concept behind the Intermediate Value Theorem. Also, we demonstrate how connectedness plays a role in applications in geographic information systems, population modeling, and motion planning in robotics. Further, we use connectedness to help distinguish topological spaces. We begin in Section 6.1 by defining connectedness and examining some aspects of it. In Section 6.2 we show that Euclidean space is connected, and then we use connectedness-related properties to distinguish between some pairs of topological spaces. In Section 6.3 we prove the Intermediate Value Theorem and then use it to derive a few interesting mathematical and applied consequences. We define path connectedness in Section 6.4. There we show that being path connected implies being connected, and we introduce a space, the topologist's whirlpool, that is connected but not path connected, establishing that connectedness and path connectedness are not equivalent properties. Finally, in Section 6.5 we explore an application of these concepts to motion planning for mobile robots.

6.1   A First Approach to Connectedness

We begin with a definition that addresses the first intuitive approach to connectedness described in the chapter introduction.

DEFINITION 6.1. *Let X be a topological space.*

   (i) *We call X **connected** if there does not exist a pair of disjoint nonempty open sets whose union is X.*

   (ii) *We call X **disconnected** if X is not connected.*

   (iii) *If X is disconnected, then a pair of disjoint nonempty open sets whose union is X is called a **separation of** X.*

EXAMPLE 6.1. Consider the two topologies on the three-point set $X = \{a, b, c\}$ in Figure 6.1. In the first topology, X is connected since there is no pair of disjoint nonempty open sets whose union equals X. However, in the second topology, X is disconnected. The pair of open sets, $U = \{a, b\}$ and $V = \{c\}$, is a separation of X.
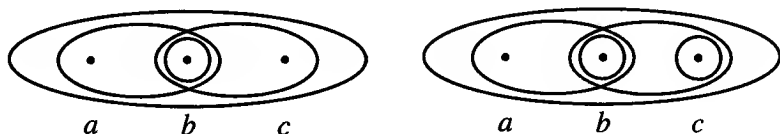
FIGURE 6.1: Two topologies on $X = \{a, b, c\}$, one connected, one not.

EXAMPLE 6.2. The subspace $X = (-1, 0) \cup (0, 1)$ of $\mathbb{R}$ is disconnected. The pair of sets, $(-1, 0)$ and $(0, 1)$, is a separation of X. (See Figure 6.2.)
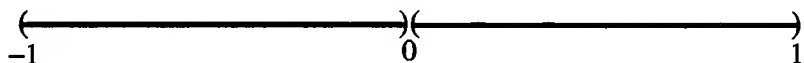
FIGURE 6.2: $X = (-1, 0) \cup (0, 1)$ is disconnected.

EXAMPLE 6.3. If a set X consists of more than one point, and it has the discrete topology, then it is disconnected. If A is any nonempty proper subset of X, then the pair of sets, A and $X - A$, is a separation of X.

EXAMPLE 6.4. A set X having the trivial topology is a connected topological space. There is no separation of X since there are no nonempty proper subsets of X that are open.

EXAMPLE 6.5. If $p \in \mathbb{R}$, then $\mathbb{R} - \{p\}$ is a disconnected topological space since the pair of sets, $U = (-\infty, p)$ and $V = (p, \infty)$, is a separation of $\mathbb{R} - \{p\}$. (See Figure 6.3.)
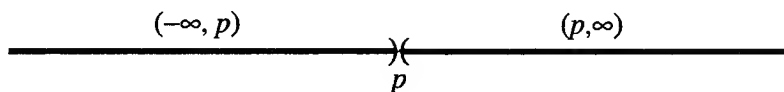
FIGURE 6.3: $\mathbb{R} - \{p\}$ is disconnected.

The following result provides an alternative formulation of connectedness:

THEOREM 6.2. *A topological space X is connected if and only if there are no nonempty proper subsets of X that are both open and closed in X.*

> *Proof.* See Exercise 6.2.   ∎

In every topological space X, the sets X and $\varnothing$ are both open and closed. Thus, Theorem 6.2 indicates that X is connected if and only if those are the only sets that are both open and closed in X.

EXAMPLE 6.6. In $\mathbb{R}$ with the lower limit topology, intervals $[a, b)$ are both open and closed. Therefore Theorem 6.2 implies that $\mathbb{R}$ is disconnected in this topology.

Our definition of connectedness applies to topological spaces. But we can extend it in a natural way to subsets of a topological space.

DEFINITION 6.3. *A set A contained in a topological space X is said to be* **connected in X** *if A is connected in the subspace topology. If A is not connected in X, we say it is* **disconnected in X**.

EXAMPLE 6.7. The subspace of $\mathbb{R}$ given by $A = [-1, 0) \cup (0, 1]$ is disconnected. The sets $U = [-1, 0)$ and $V = (0, 1]$ form a separation of A. Hence A is disconnected in $\mathbb{R}$.

The next theorem yields an alternate characterization of disconnected sets in a topological space X.

THEOREM 6.4. *A set A is disconnected in X if and only if there exist open sets U and V in X such that $A \subset U \cup V$, $U \cap A \neq \varnothing$, $V \cap A \neq \varnothing$, and $U \cap V \cap A = \varnothing$.*

> *Proof.* Suppose that A is disconnected in X. Then there exist nonempty sets P and Q that are open in A, disjoint, and such that $P \cup Q = A$. Since P and Q are open in A there exist sets U and V that are open in X and such that $U \cap A = P$ and $V \cap A = Q$. Clearly, $A \subset U \cup V$, $U \cap A \neq \varnothing$, $V \cap A \neq \varnothing$, and $U \cap V \cap A = \varnothing$.
>
> Now suppose that U and V are open sets in X such that $A \subset U \cup V$, $U \cap A \neq \varnothing$, $V \cap A \neq \varnothing$, and $U \cap V \cap A = \varnothing$. If we let $P = U \cap A$ and $Q = V \cap A$, then it follows that the pair of sets, P and Q, is a separation of A in the subspace topology, and therefore A is disconnected in X.   ∎

Sets such as U and V in Theorem 6.4 provide us with another notion of separation, defined as follows:

DEFINITION 6.5. *Let A be a subspace of a topological space X. If U and V are open sets in X such that $A \subset U \cup V$, $U \cap A \neq \varnothing$, $V \cap A \neq \varnothing$, and $U \cap V \cap A = \varnothing$, then we say that the pair of sets, U and V, is a **separation** of A in X. (See Figure 6.4.)*
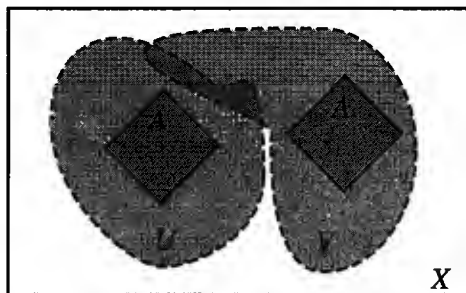
FIGURE 6.4: The sets U and V form a separation of A in X.

IMPORTANT NOTE: *For a pair of sets U and V to be a separation of A in X, we do not require that $U \cap V$ is empty; we only need $U \cap V$ to be disjoint from A.*

EXAMPLE 6.8. Let A be the subset of the plane consisting of the two curves $y = e^x$ and $y = 0$, as illustrated in Figure 6.5. Is A connected? No. We can form a separation of A in the plane by letting U and V be the sets of points on either side of the graph of $y = e^{x-1}$, shown as the dashed curve in the figure. Thus A is disconnected in the plane.

Another separation of A in the plane is given by letting U' be the set of points below the curve $y = e^x$ and V' be the set of points above the x-axis. Note that U' and V' are not disjoint, but they still form a separation of A in the plane since their intersection is disjoint from A.
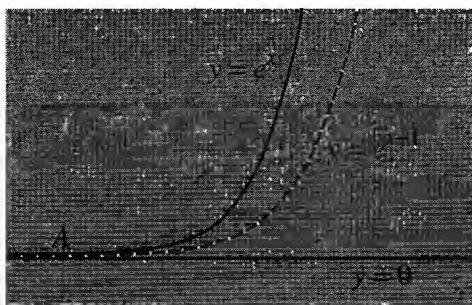
FIGURE 6.5: The set A is disconnected.

It is apparent that connectedness is a topological property since it is defined only in terms of open sets. It is straightforward to show that if X and Y are homeomorphic, then X is connected if and only if Y is connected. However, as the following theorem indicates, we do not need the full strength of a homeomorphism to preserve connectedness:

THEOREM 6.6. *If X is connected and $f : X \to Y$ is continuous, then $f(X)$ is connected in Y.*

> **Proof.** Suppose that $f(X)$ is not connected in Y. Then there exist open sets U and V that form a separation of $f(X)$ in Y. The function f is continuous, and therefore $f^{-1}(U)$ and $f^{-1}(V)$ are open in X. Both U and V have nonempty intersection with $f(X)$; thus $f^{-1}(U)$ and $f^{-1}(V)$ are nonempty. Furthermore $f(X) \subset U \cup V$, implying that $X \subset f^{-1}(U) \cup f^{-1}(V)$. Finally, since $U \cap V \cap f(X) = \varnothing$, it follows that $f^{-1}(U)$ and $f^{-1}(V)$ are disjoint. Therefore the pair of sets, $f^{-1}(U)$ and $f^{-1}(V)$, is a separation of X, contradicting the assumption that X is connected. Hence, $f(X)$ is connected in Y. ∎

The following lemma will be useful to us in a few proofs in this chapter:

LEMMA 6.7. *Let C and D be subsets of a topological space X. Assume that C is connected and $C \subset D$. Further assume that U and V form a separation of D in X. Then either $C \subset U$ or $C \subset V$.*

> **Proof.** Suppose that neither $C \subset U$ nor $C \subset V$. Then $U \cap C \neq \varnothing$ and $V \cap C \neq \varnothing$. It follows that U and V form a separation of C in X, contradicting the assumption that C is connected. ∎

Our next theorem indicates that if C is connected in X, and we add limit points to C, then the resulting set is also connected in X.

THEOREM 6.8. *Let C be connected in X, and assume that $C \subset A \subset Cl(C)$. Then A is connected in X.*

> **Proof.** Suppose that A is not connected in X, and let U and V form a separation of A in X. Then by Lemma 6.7, either $C \subset U$ or $C \subset V$. We may assume, without loss of generality, that $C \subset U$. Hence $C \cap V = \varnothing$. But, since U and V form a separation of A in X, it follows that $A \cap V \neq \varnothing$. Pick $x \in A \cap V$. Now, $x \in A$ and $A \subset Cl(C)$ imply $x \in Cl(C)$. But $x \in V$, an open set in X which is disjoint from C. So x cannot be in the closure of C, yielding a contradiction. Thus, it follows that A is connected in X. ∎

A union of connected subsets of a topological space is not necessarily connected. For example, $A = \{0\}$ and $B = \{1\}$ are connected sets in $\mathbb{R}$ with the standard topology, but $A \cup B$ is not connected in $\mathbb{R}$. However, as the following theorem indicates, if the sets in a collection of connected subsets of a topological space have at least one point in common, then we are ensured that the union of the subsets is connected:

THEOREM 6.9. *Let X be a topological space, and let $\{C_\alpha\}_{\alpha \in A}$ be a collection of connected subsets of X such that $\bigcap_{\alpha \in A} C_\alpha \neq \varnothing$. Then $\bigcup_{\alpha \in A} C_\alpha$ is connected in X.*

> ***Proof.*** Suppose that $\bigcup_{\alpha \in A} C_\alpha$ is not connected in X. Thus there exist sets U and V that form a separation of $\bigcup_{\alpha \in A} C_\alpha$ in X. Let x be in $\bigcap_{\alpha \in A} C_\alpha$. Then either $x \in U$ or $x \in V$, but both cannot hold. We may assume, without loss of generality, that x lies in U and does not lie in V. Lemma 6.7 implies that for all $\alpha \in A$, either $C_\alpha \subset U$ or $C_\alpha \subset V$. Since $x \in U$ and $x \notin V$, it follows that $C_\alpha \subset U$ for all $\alpha \in A$. Thus $\bigcup_{\alpha \in A} C_\alpha \subset U$, contradicting the assumption that U and V form a separation of $\bigcup_{\alpha \in A} C_\alpha$ in X. Therefore $\bigcup_{\alpha \in A} C_\alpha$ is connected in X. ∎

While Theorem 6.9 is about unions of connected spaces, we can use it to prove that a product of connected spaces is connected. We do that in the following theorem:

THEOREM 6.10. *Let $X_1, \dots, X_n$ be connected spaces. Then the product space $X_1 \times \dots \times X_n$ is connected.*

> ***Proof.*** We prove the result for a product of two spaces. The general result can then be shown by induction. Assume that X and Y are connected topological spaces. We prove that $X \times Y$ is connected. First, note that for every $x \in X$, the subspace $\{x\} \times Y$ of $X \times Y$ is homeomorphic to Y and is therefore connected. Similarly, for every $y \in Y$, the subspace $X \times \{y\}$ of $X \times Y$ is connected. Thus, by Theorem 6.9, for every $x \in X$ and $y \in Y$ the set $(\{x\} \times Y) \cup (X \times \{y\})$ is connected in $X \times Y$. (See Figure 6.6.)
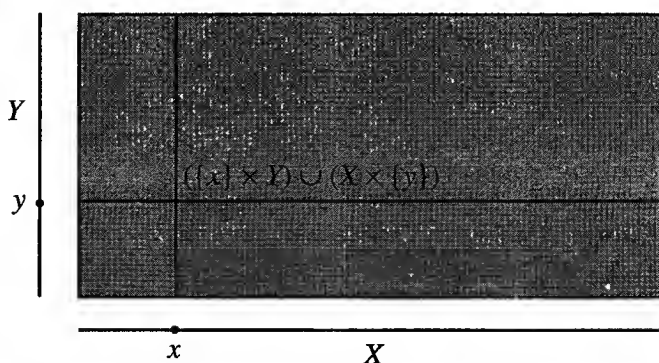
FIGURE 6.6: The set $(\{x\} \times Y) \cup (X \times \{y\})$ is connected in $X \times Y$.

Now fix $x_0 \in X$ and let y vary. Each set $(\{x_0\} \times Y) \cup (X \times \{y\})$ contains the set $\{x_0\} \times Y$. It then follows by Theorem 6.9 that

$$\bigcup_{y \in Y} ((\{x_0\} \times Y) \cup (X \times \{y\}))$$

is connected in $X \times Y$. Furthermore,

$$\bigcup_{y \in Y} ((\{x_0\} \times Y) \cup (X \times \{y\})) = X \times Y,$$

implying that $X \times Y$ is connected. (See Figure 6.7.)
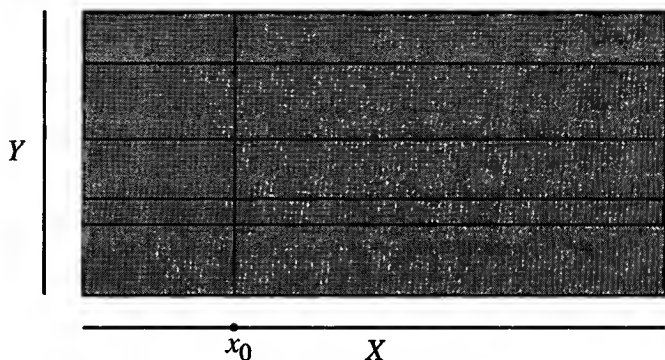
FIGURE 6.7: The product $X \times Y$ is the union, over y in Y, of the subspaces $(\{x_0\} \times Y) \cup (X \times \{y\})$.

∎

A topological space can naturally be partitioned into a collection consisting of its largest connected subsets. We formalize this idea in what follows.

Let X be a topological space. Define a relation $\sim_C$ on X by $x \sim_C y$ if x and y lie in a connected subset of X. We claim that $\sim_C$ is an equivalence relation. It is clear that $x \sim_C x$ for every $x \in X$ and that $x \sim_C y$ implies $y \sim_C x$ for every $x, y \in X$. Suppose x, y, and z in X are such that $x \sim_C y$ and $y \sim_C z$. Then x and y lie in a connected set C in X, and y and z lie in a connected set C' in X. Both C and C' contain the point y. Therefore $C \cup C'$ is connected by Theorem 6.9. Since x and z lie in $C \cup C'$, it follows that $x \sim_C z$. Thus, for all $x, y, z \in X$, if $x \sim_C y$ and $y \sim_C z$, then $x \sim_C z$. Hence $\sim_C$ is an equivalence relation.

DEFINITION 6.11. *Let X be a topological space and let $\sim_C$ be the equivalence relation on X defined by $x \sim_C y$ if x and y lie in a connected subset of X. The **components** of X are the equivalence classes of the equivalence relation $\sim_C$.*

Since the components of a topological space X are equivalence classes under an equivalence relation, they form a partition of X. That components are the largest connected subsets of X follows from the first two parts of the next theorem.

THEOREM 6.12. *Let X be a topological space.*

(i) Each component of X is connected in X.

(ii) *If A is connected in X, then A is a subset of a component of X.*

(iii) *Each component of X is a closed subset of X.*

We prove (i) here. For proofs of (ii) and (iii), see Exercise 6.9.

> **Proof of (i).** Let X be a topological space and C be a component of X. We prove that C is connected. Pick a point $p \in C$. For every $x \in C$, $x \sim_C p$ since C is an equivalence class under the equivalence relation $\sim_C$ that determines the components. Therefore, by the definition of $\sim_C$, there exists a connected set C_x containing x and p.
>
> We claim that $C_x \subset C$. To prove the claim, suppose $y \in C_x$. Then y and p both lie in the connected set C_x, implying that $y \sim_C p$. Therefore y is in the same equivalence class as p, namely C, and it follows that $C_x \subset C$.
>
> Now, by the Union Lemma we have $\bigcup_{x \in C} C_x = C$. Therefore C is the union of the connected sets C_x, and each of these connected sets contains the point p. By Theorem 6.9, it follows that C is connected. ∎

EXAMPLE 6.9. Consider $\mathbb{R}$ with the finite complement topology. It is a connected topological space. (See Exercise 6.1.) Therefore, in this topology there is only one component, the whole space $\mathbb{R}$.

EXAMPLE 6.10. Consider the real line $\mathbb{R}$ with the lower limit topology. Here we have a situation that is very different from the one in Example 6.9. We claim that the components are the one-point subsets of $\mathbb{R}$. To prove this, we show that if a subset of $\mathbb{R}$ consists of more than one point, then it is disconnected. Thus let A be a subset of $\mathbb{R}$ containing two points $x < y$. Pick z such that $x < z < y$. The pair of sets, $(-\infty, z)$ and $[z, \infty)$, is a separation of A in $\mathbb{R}$ with the lower limit topology, and therefore A is disconnected. Thus, the only nonempty connected subsets of $\mathbb{R}$ in this topology are the one-point sets, and it follows that they are the components.

DEFINITION 6.13. *A topological space X is called **totally disconnected** if the components of X are the one-point subsets of X.*

By Example 6.10, $\mathbb{R}$ with the lower limit topology is totally disconnected. Another totally disconnected topological space is the set of rational numbers $\mathbb{Q}$ in the standard topology. (See Exercise 6.7.)

The following theorem indicates that homeomorphisms map components to components:

THEOREM 6.14. *Let $f : X \to Y$ be a homeomorphism. If C is a component of X, then $f(C)$ is a component of Y.*

Proof. Suppose that $f : X \to Y$ is a homeomorphism and C is a component of X. Then C is connected in X by Theorem 6.12. Thus Theorem 6.6 implies that $f(C)$ is connected in Y. Again by Theorem 6.12, it follows that $f(C)$ is a subset of a component D of Y. We claim that $f(C) = D$, to complete the proof of the theorem.

The component D is connected in Y, and f^{-1} is a continuous function; therefore $f^{-1}(D)$ is connected in X. Furthermore, $f(C) \subset D$ implies that $C \subset f^{-1}(D)$. But C is a component of X, and C is a subset of the connected set $f^{-1}(D)$. Therefore $C = f^{-1}(D)$. Hence, $f(C) = D$, as we wished to show.   ∎

Exercises for Section 6.1

6.1. Prove that an infinite set with the finite complement topology is a connected topological space.

6.2. **Prove Theorem 6.2**: A topological space X is connected if and only if there are no nonempty proper subsets of X that are both open and closed in X.

6.3. Prove that a topological space X is connected if and only if every nonempty proper subset of X has a nonempty boundary.

6.4. (a) Prove that for all integers $n \geq 0$ the set $\{-n, -n+1, \ldots, n-1, n\}$ is connected in the digital line.
   (b) Use part (a) to prove that the digital line is connected.

6.5. In the topology generated by the basis $\mathcal{B} = \{(-a, a) \mid a \in \mathbb{R}\}$, show that $\mathbb{R}$ is connected.

6.6. Let X be a set and assume $p \in X$. Prove that every subset of X is connected in the particular point topology PPX_p on X and in the excluded point topology EPX_p on X. (See Exercises 1.7 and 1.8.)

6.7. (a) Prove that if a topological space X has the discrete topology, then X is totally disconnected.
   (b) Let $\mathbb{Q}$ be the set of rational numbers with the standard topology. Prove that $\mathbb{Q}$ is totally disconnected. (This exercise and Example 6.9 demonstrate that the converse to the result in part (a) does not hold. In both cases, the space is totally disconnected but does not have the discrete topology.)

6.8. Let X be a topological space and $\{A_j\}_{j \in \mathbb{Z}_+}$ be a collection of connected subsets of X such that $A_j \cap A_{j+1} \neq \varnothing$ for each $j \geq 1$. Prove that $\bigcup_{j \in \mathbb{Z}_+} A_j$ is connected in X.

6.9. **Prove Theorem 6.12, parts (ii) and (iii)**: Let X be a topological space.
   (a) If A is connected in X, then A is a subset of a component of X.
   (b) Each component of X is a closed subset of X.
   (c) Provide an example showing that the components of X are not necessarily open subsets of X.

6.10. The following examples demonstrate that the condition $U \cap V \cap A = \varnothing$ is appropriate in the definition of a separation of A in X and that the condition would be too strong if it required $U \cap V = \varnothing$:
   (a) Find an example of a topology on $X = \{a, b, c\}$ and a disconnected subset A such that every pair of sets, U and V, that is a separation of A in X satisfies $U \cap V \cap (X - A) \neq \varnothing$.
   (b) Find a topology on $\mathbb{R}$ and a disconnected subset A such that every pair of sets, U and V, that is a separation of A in $\mathbb{R}$ satisfies $U \cap V \cap (\mathbb{R} - A) \neq \varnothing$.

Supplementary Exercises: GIS Revisited

In the following exercises, we continue the investigation of the intersection value and its application to geographic information systems, which we introduced in Section 2.4. Recall that in this application we are particularly interested in regularly closed planar sets, the interiors of which are not the union of two disjoint nonempty open sets. With the concept of connectedness available to us, we can now define these sets of interest in a more concise manner. Furthermore, we work here with a more general collection of sets than the planar spatial regions introduced in Section 2.4. In particular, we do not require the sets to be subsets of $\mathbb{R}^2$.

DEFINITION 6.15. *Given a topological space X, a* **spatial region** *A in X is a nonempty, regularly closed, proper subset of X that has a connected interior.*

SE 6.11. Let X be a topological space. Prove that if A is a proper subset of X, and A is the closure of a nonempty, connected open set, then A is a spatial region. (Note: The interior of A need not be the open set referred to in the statement of the problem, but it can be shown that the interior of A is connected as a result of the fact that A is the closure of a connected open set.)

Recall the definition of the intersection value, given in Section 2.4. Since spatial regions are regularly closed sets, Corollary 2.19 applies, and it follows that the intersection values $(1, 0, 1, 0)$, $(0, 0, 1, 0)$, $(1, 0, 0, 1)$, $(0, 0, 0, 1)$, $(1, 0, 1, 1)$, and $(0, 0, 1, 1)$ are not possible for a pair of spatial regions. The following exercise eliminates one more possible intersection value, as long as we assume that we have two spatial regions in a connected topological space. The plane is connected (see Corollary 6.18), and therefore this exercise establishes the assertion in Theorem 2.21 that the only intersection values that are possible for a pair of planar spatial regions are $(1, 1, 1, 0)$, $(0, 1, 1, 0)$, $(1, 1, 0, 1)$, $(0, 1, 0, 1)$, $(1, 1, 0, 0)$, $(0, 0, 0, 0)$, $(1, 0, 0, 0)$, $(0, 1, 1, 1)$, and $(1, 1, 1, 1)$.

SE 6.12. Let X be a connected topological space, and assume that A and B are spatial regions in X. Then the intersection value for A and B in X cannot equal $(0, 1, 0, 0)$. (Hint: Prove that $A \cap B$ is a nonempty set that is both open and closed, and then apply Theorem 6.2.)

Consider Theorem 2.21 again. We addressed parts of its the proof in Exercises 2.30 and SE 6.12. The rest of Theorem 2.21 follows by showing that, if the intersection value for spatial regions A and B in a connected topological space X is as depicted in the left column in the following table, then the relationship between A and B is as depicted in the corresponding entry in the right column. We address these results in the subsequent exercises.

Intersection Value	Relationship
$(1, 1, 1, 0)$	$A \subset B$
$(0, 1, 1, 0)$	$A \subset \text{Int}(B)$
$(1, 1, 0, 1)$	$B \subset A$
$(0, 1, 0, 1)$	$B \subset \text{Int}(A)$
$(1, 1, 0, 0)$ ·	$A = B$

The following lemma will be helpful:

LEMMA 6.16. *Let X be a connected topological space. Assume that $C, D \subset X$ and that C is connected. If $C \cap \partial D = \varnothing$, then either $C \subset Int(D)$ or $C \cap Cl(D) = \varnothing$.*

SE 6.13. Prove Lemma 6.16.

The following three exercises address the first, second, and fifth rows in the foregoing table. Lemma 6.16 should provide some assistance. The proofs for the third and forth rows are similar to those for the first and second rows.

SE 6.14. Let A and B be spatial regions in a connected topological space X. Prove that if the intersection value for A and B in X is $(1,1,1,0)$, then $A \subset B$.

SE 6.15. Let A and B be spatial regions in a connected topological space X. Prove that if the intersection value for A and B in X is $(0,1,1,0)$, then $A \subset Int(B)$.

SE 6.16. Let A and B be spatial regions in a connected topological space X. Prove that if the intersection value for A and B in X is $(1,1,0,0)$, then $A = B$.

6.2   *Distinguishing Topological Spaces via Connectedness*

In this section, we prove that Euclidean n-space in the standard topology is a connected topological space. We first show that the real line in the standard topology is connected. The general result for $\mathbb{R}^n$ then follows by Theorem 6.10. We also show how the concept of connectedness can be used to help distinguish between topological spaces (that is, determine that particular pairs of topological spaces are not homeomorphic).

To prove that the real line is connected, we use the following properties of the real number system:

  (i) The Least Upper Bound Property: Every subset of $\mathbb{R}$ that is bounded above has a least upper bound.

  (ii) If $x, y \in \mathbb{R}$ with $x < y$ then there exists z such that $x < z < y$.

THEOREM 6.17. *The real line $\mathbb{R}$ in the standard topology is a connected topological subspace.*

  Proof. Suppose it is not, and let U and V form a separation of $\mathbb{R}$. Pick $u \in U$ and $v \in V$. We may assume $u < v$ without loss of generality. Let $U' = U \cap [u, v]$ and $V' = V \cap [u, v]$. (See Figure 6.8.) It follows that $U' \cup V' = [u, v]$. Since U' is bounded from above (by v), U' has a least upper bound; call it c. We have $u \leq c \leq v$. We derive a contradiction by showing that $c \notin U'$ and $c \notin V'$.

  To show that $c \notin V'$, we assume that $c \in V'$ and derive a contradiction. Since $u \notin V'$ and V' is open in $[u, v]$, it follows that there exists d such that $u < d < c$ and $(d, c] \subset V'$. This implies that d is an upper
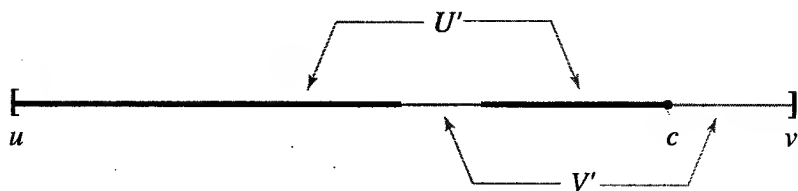
FIGURE 6.8: The sets U' and V' in $[u, v]$.

bound of U' and that d is less than the least upper bound c. This is a contradiction and thus $c \notin V'$.

Next we show that $c \notin U'$. We do this by contradiction as well. Therefore, assume that $c \in U'$. Since U' is open in $[u, v]$ and $v \notin U'$, there exists d such that $[c, d) \subset U'$. For any $e \in (c, d)$ it follows that $e \in U'$ and $e > c$, contradicting the fact that c is an upper bound of U'. Thus $c \notin U'$.

Therefore $c \notin U'$ and $c \notin V'$. But $c \in [u, v]$ and $U' \cup V' = [u, v]$. With this final contradiction it now follows that $\mathbb{R}$ in the standard topology is a connected topological space.   ∎

Theorems 6.10 and 6.17 imply the following corollary:

COROLLARY 6.18. *Euclidean n-space $\mathbb{R}^n$ is a connected topological space.*

EXAMPLE 6.11. All interval subsets of $\mathbb{R}$ are connected in $\mathbb{R}$. Why? Open intervals (a, b), $(-\infty, b)$, and (a, ∞) are connected because they are homeomorphic to $\mathbb{R}$. Theorem 6.8 then implies that all other interval subsets of $\mathbb{R}$ are connected because they can be obtained from open intervals by adding limit points.

EXAMPLE 6.12. The n-sphere S^n is a connected topological space. We prove this for the sphere S^2, but a similar approach works to prove that S^n is connected.

The stereographic projection function introduced in Example 4.16 is a homeomorphism between $\mathbb{R}^2$ and the subspace $S^2 - \{N\}$ of S^2 that is obtained by removing the north pole N. Since $\mathbb{R}^2$ is connected, so is $S^2 - \{N\}$. Thus by Theorem 6.8, the closure of $S^2 - \{N\}$ in S^2 is connected as well. That closure is, of course, the whole space S^2, and therefore the sphere S^2 is a connected topological space.

EXAMPLE 6.13. Since the interval $I = [0, 1]$ is connected, so is the square $I \times I$. The annulus, Möbius band, torus, Klein bottle, sphere, and projective plane are all either obtained from a square by gluing pairs of edges or

homeomorphic to such a space. (See Figure 6.9.) Therefore there are quotient maps from the square onto each of these spaces, and these maps are continuous functions with a connected domain. It follows that each of these spaces is connected.
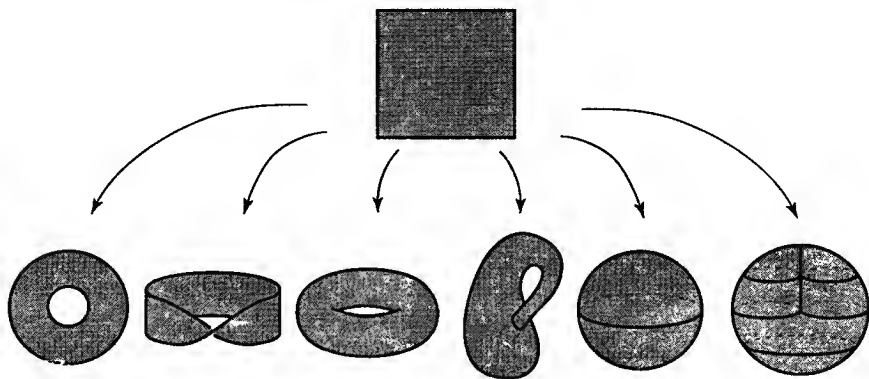
FIGURE 6.9: Quotient spaces of the square are connected.

EXAMPLE 6.14. The topological space $\mathbb{R}^2 - \{O\}$ is called the **punctured plane** (where $O = (0, 0)$ is the origin in the plane). We show that it is connected. Each open half plane $H_+ = \{(x, y) \mid y > 0\}$ and $H_- = \{(x, y) \mid y < 0\}$ is homeomorphic to $\mathbb{R}^2$. (See Example 4.14.) Therefore H_+ and H_- are both connected in $\mathbb{R}^2$. By Theorem 6.8, we can add limit points to each and maintain the fact that they are connected. So to both H_+ and H_- we add all of the x-axis except for the origin. The resulting sets are also connected in $\mathbb{R}^2$. Since the resulting sets intersect, their union is also connected, and their union is exactly $\mathbb{R}^2 - \{O\}$.

Note that if p is a point in $\mathbb{R}^2$, then $\mathbb{R}^2 - \{p\}$ is connected since $\mathbb{R}^2 - \{p\}$ is homeomorphic to $\mathbb{R}^2 - \{O\}$.

Now that we have a collection of topological spaces that are connected, we demonstrate how connectedness can be used to distinguish between topological spaces. A line is not a circle is not a plane is not a sphere. How can we prove this? We will see that connectedness helps.

As we indicated in Example 6.5, the removal of a point from the real line results in a disconnected topological space. This is an important enough notion that we provide a specific definition:

DEFINITION 6.19. *Let X be a connected topological space. A **cutset** of X is a subset S of X such that $X - S$ is disconnected. A **cutpoint** of X is a point $p \in X$ such that $\{p\}$ is a cutset of X. A cutset or cutpoint of X is said to separate X.*

EXAMPLE 6.15. The plane $\mathbb{R}^2$ is connected. If we remove the circle S^1, we are left with two disjoint nonempty open sets as shown in Figure 6.10. Hence, $\mathbb{R}^2 - S^1$ is disconnected, implying that S^1 is a cutset of the plane.

In fact, the Jordan Curve Theorem, a classic theorem in topology, asserts that every simple closed curve in the plane is a cutset of the plane. This general result is much more difficult to establish than the specific case of the circle S^1. We prove the Jordan Curve Theorem in Chapter 11.
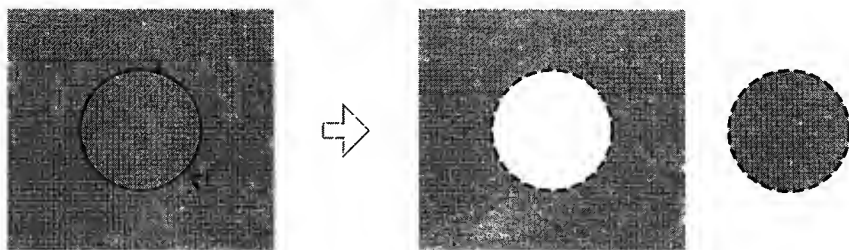
FIGURE 6.10: The circle S^1 is a cutset of the plane $\mathbb{R}^2$.

As we might expect from the situation depicted in Figure 6.11, if X is a connected topological space, then the boundary of a subset A of X is a cutset of X if both $\text{Int}(A)$ and $\text{Int}(X - A)$ are nonempty. (See Exercise 6.23.)
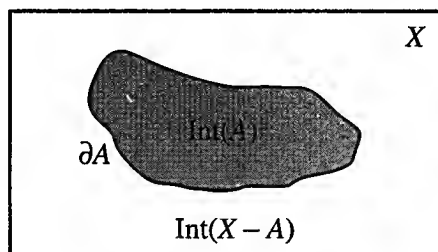
FIGURE 6.11: The boundary of the set A is a cutset of X.

The following theorem establishes that cutsets and cutpoints are preserved under homeomorphism. Hence we can use them to distinguish between topological spaces.

THEOREM 6.20. *Let $f : X \to Y$ be a homeomorphism. If S is a cutset of X, then $f(S)$ is a cutset of Y.*

Proof. See Exercise 6.24.   ∎

It follows immediately from Theorem 6.20 that if $f : X \to Y$ is a homeomorphism and p is a cutpoint of X, then $f(p)$ is a cutpoint of Y.

EXAMPLE 6.16. Every point p in $\mathbb{R}$ is a cutpoint. In contrast, no point q in the plane $\mathbb{R}^2$ is a cutpoint since $\mathbb{R}^2 - \{q\}$ is connected for all $q \in \mathbb{R}^2$. (See Figure 6.12.) Thus the line is not homeomorphic to the plane.
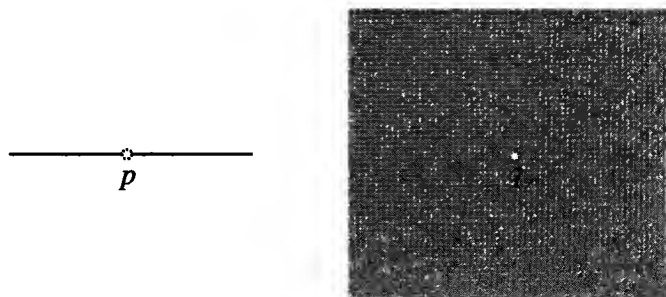
FIGURE 6.12: $\mathbb{R} - \{p\}$ is disconnected, but $\mathbb{R}^2 - \{q\}$ is connected.

EXAMPLE 6.17. For every $q \in S^1$, the space $S^1 - \{q\}$ is homeomorphic to $\mathbb{R}$. Since $\mathbb{R}$ is connected, $S^1 - \{q\}$ is connected as well. Therefore no point in the circle S^1 is a cutpoint. (See Figure 6.13.) But since every point in the line is a cutpoint, it follows that the circle is not homeomorphic to the line. Furthermore, for every $r \in S^2$, the space $S^2 - \{r\}$ is homeomorphic to the plane and therefore is connected. Thus no point in the sphere S^2 is a cutpoint, implying that the sphere is not homeomorphic to the line.
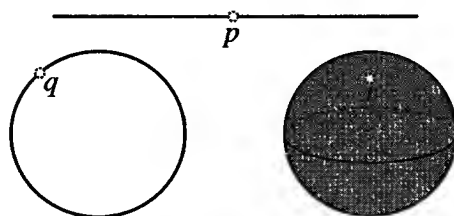
FIGURE 6.13: $\mathbb{R} - \{p\}$ is disconnected, but $S^1 - \{q\}$ and $S^2 - \{r\}$ are connected.

EXAMPLE 6.18. Every two-point set $\{p, p'\}$ in the circle separates the circle into two open intervals, and therefore every two-point set is a cutset of the circle. (See Figure 6.14.) However, if we remove a two-point set $\{q, q'\}$ from the sphere, we obtain a space that is homeomorphic to the punctured plane, $\mathbb{R}^2 - \{O\}$. Since $\mathbb{R}^2 - \{O\}$ is connected, it follows that no two-point set in the sphere is a cutset. Thus, the circle is not homeomorphic to the sphere.

FIGURE 6.14: $S^1 - \{p, p'\}$ is disconnected, but $S^2 - \{q, q'\}$ is connected.

So far we have distinguished the line from the plane, circle, and sphere, and we have distinguished the circle from the sphere. We can similarly distinguish the circle from the plane. (See Exercise 6.29.) However, when we consider the sphere and the plane, we find that no one-point set, two-point set, or even countably infinite set separates either. (See Exercise 6.43.) Therefore, we cannot distinguish the sphere from the plane by removing simple sets, as in the previous examples. We show in 'Chapter 7 that the topological property of compactness enables us to distinguish between these two spaces. (See Exercise 7.17.)

Exercises for Section 6.2

6.17. **(a)** Show that open balls and closed balls in the standard metric on $\mathbb{R}^n$ are connected in $\mathbb{R}^n$.

(b) Prove that each of the sets, A, B, and C, appearing in Figure 6.15, is a connected subset of the plane. (The set C consists of eight open balls and the eight points of tangency between them, as shown.)
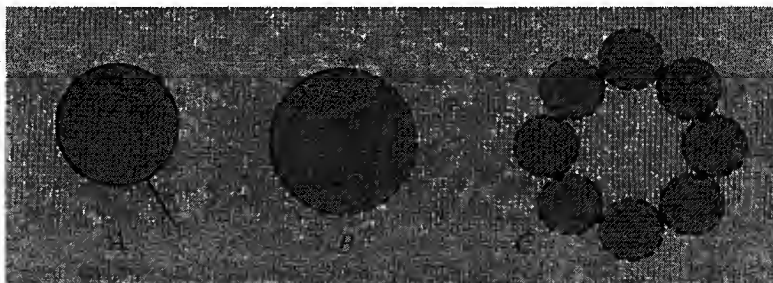
FIGURE 6.15: Prove that each of these subsets of the plane is connected.

6.18. Give examples of subsets A and B in $\mathbb{R}^2$ such that

(a) A and B are connected, but $A \cap B$ is not.

(b) A and B are connected, but $A - B$ is not.

(c) A is connected, B is disconnected, and $A \cap B$ is connected.

(d) A and B are disconnected, but $A \cup B$ is connected.

(e) A and B are connected, $\mathrm{Cl}(A) \cap \mathrm{Cl}(B) \neq \varnothing$, and $A \cup B$ is disconnected.

6.19. Let $B \subset \mathbb{R}^n$ be bounded in the Euclidean metric. Prove that the complement of B in $\mathbb{R}^n$ has exactly one unbounded component.

6.20. In each of the following cases, prove whether or not the given set C is a cutset of the connected topological space X:

(a) $C = \{b\}$ and $X = \{a, b, c\}$ with topology $\{\varnothing, \{b\}, \{a, b\}, \{b, c\}, X\}$.

(b) $C = \{c\}$ and X is the same as in (a).

(c) $C = \{0\}$ and $X = PP\mathbb{R}_0$, the particular point topology on $\mathbb{R}$ with the origin as the particular point.

(d) $C = \{-1, 1\}$ and $X = \mathbb{R}_{fc}$, the real line in the finite complement topology.

(e) $C =$ the circle S^1 and $X = \mathbb{R}^2$.

(f) $C =$ the core curve in the Möbius band X, as shown in Figure 6.16.
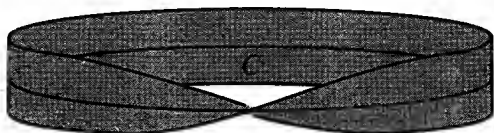
FIGURE 6.16: Is $X - C$ connected?

6.21. Prove that each $n \in \mathbb{Z}$ is a cutpoint of the digital line.

6.22. Can you cut a Klein bottle into two Möbius bands? Find a cutset C for the Klein bottle K such that

> **(i)** C is a simple closed curve in K, and
>
> **(ii)** $K - C$ is a union of two disjoint open sets such that the closure of each in K is homeomorphic to a Möbius band.

6.23. Let X be a connected topological space and A be a subset of X. Prove that if $\text{Int}(A)$ and $\text{Int}(X - A)$ are nonempty, then ∂A is a cutset, and the pair of sets, $\text{Int}(A)$ and $\text{Int}(X - A)$, is a separation of $X - \partial A$.

6.24. Prove Theorem 6.20: Let $f : X \to Y$ be a homeomorphism. If S is a cutset of X, then $f(S)$ is a cutset of Y.

6.25. Prove that for every $n \geq 2$ the line is not homeomorphic to $\mathbb{R}^n$.

6.26. Prove that for every $n \geq 2$ neither the line nor the circle is homeomorphic to S^n.

6.27. Consider $[0, 1]$, $[0, 1)$, and $(0, 1)$ as subspaces of $\mathbb{R}$ in the standard topology. Prove that no two of these spaces are homeomorphic to each other. (This result establishes the claim made in Example 4.12 that, given the three collections of intervals defined in the example, there is no homeomorphism from an interval in one collection to an interval in another.)

6.28. (a) Prove that the interval $[0, 2\pi)$ is not homeomorphic to the circle S^1. (This exercise establishes the claim made in Example 4.13 that there is no homeomorphism between these two spaces.)

(b) Prove that no interval subspace of $\mathbb{R}$ is homeomorphic to the circle S^1.

6.29. Prove that no two-point set separates the plane, and use that result to argue that the circle is not homeomorphic to the plane.

6.30. Consider the following definition:

> **DEFINITION 6.21.** *Let X be a connected topological space, and assume that $n \in \mathbb{Z}_+$ with $n \geq 2$. A point p in X is said to be a **local cutpoint of order n in** X if there is a connected neighborhood U of p such that $U - \{p\}$ has n components.*

> Prove that if $f : X \to Y$ is a homeomorphism and p is a local cutpoint of order n in X, then $f(p)$ is a local cutpoint of order n in Y.

6.31. Begin with a collection of topological graph representations of the letters of the alphabet, grouped into equivalence classes by topological equivalence. (See Exercise 4.37.) Use Exercise 6.30 to verify that if two representations are in different equivalence classes in your grouping, they are not homeomorphic.

6.3   The Intermediate Value Theorem

One of the most important theorems in calculus is the Intermediate Value Theorem. It indicates that if a continuous function maps a closed interval $[a, b]$ to the real line $\mathbb{R}$, then the function takes on every value between the values $f(a)$ and $f(b)$.

THEOREM 6.22. *The Intermediate Value Theorem on [a, b].* *Let* $f : [a, b] \to \mathbb{R}$ *be continuous, and assume that s lies between* $f(a)$ *and* $f(b)$. *Then there exists at least one* $c \in [a, b]$ *such that* $f(c) = s$. *(See Figure 6.17.)*
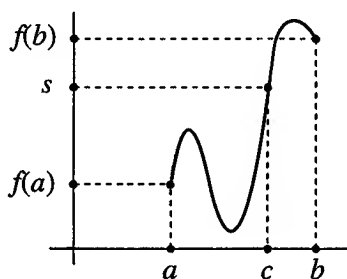
FIGURE 6.17: There exists $c \in [a, b]$ such that $f(c) = s$.

The Intermediate Value Theorem on $[a, b]$ follows as a consequence of a more general version (Theorem 6.24) that we will subsequently prove.

EXAMPLE 6.19. Weighing 320 pounds on January 1, 2001, Jared made a New Year's resolution to change his lifestyle in order to lose some weight. So he decided to stop having lunch at his favorite burger joint, and he started frequenting the sandwich shop across the street. Amazingly, Jared lost 50 pounds over the next 30 weeks. If we let $W(t)$ be Jared's weight as a function of time t, in weeks, since he started his diet, then $W(0) = 320$ and $W(30) = 270$. For every weight between 270 and 320, there was a time when Jared weighed that much. In particular, there was a time when he weighed 300 pounds. In fact, he may have weighed 300 pounds a few times, but the Intermediate Value Theorem tells us only that he weighed exactly that much at least once over the 30-week time interval.

An example of the utility of the Intermediate Value Theorem is given in the following corollary:

COROLLARY 6.23. *Let $f : [a, b] \to \mathbb{R}$ be a continuous function such that $f(a)$ and $f(b)$ have opposite signs. Then the equation $f(x) = 0$ has a solution between a and b.*

Corollary 6.23 assists us in finding solutions to an equation in the form $f(x) = 0$. For example, if f is a continuous function on $\mathbb{R}$ and we find an interval where f is negative at one end and positive at the other, then Corollary 6.23 guarantees that there is a solution to the equation $f(x) = 0$ in the interval.

Although the Intermediate Value Theorem on $[a, b]$ is first encountered in calculus, the theorem can be stated in much greater generality. It then becomes apparent that, in fact, this is a theorem from topology.

THEOREM 6.24. *The Intermediate Value Theorem (General Version).   Let X be a connected topological space and $f : X \to \mathbb{R}$ be continuous. If $p, q \in f(X)$ and $p \le r \le q$, then $r \in f(X)$.*

> **Proof.** Suppose $f : X \to \mathbb{R}$ is continuous, $p, q \in f(X)$, and $p \le r \le q$. If $r = p$ or $r = q$, then we immediately have that $r \in f(X)$. Therefore, we only need to consider the case where $p < r < q$. Note that $f(X)$ is connected in $\mathbb{R}$ since X is connected and f is continuous. We prove by contradiction that $r \in f(X)$. Thus, suppose that $r \notin f(X)$. Then $U = (-\infty, r)$ and $V = (r, \infty)$ are disjoint open subsets of $\mathbb{R}$ whose union contains $f(X)$. Since $p \in U$ and $q \in V$, it follows that $f(X)$ intersects both U and V. Hence, U and V form a separation of $f(X)$ in $\mathbb{R}$, as illustrated in Figure 6.18. But this contradicts the fact that $f(X)$ is connected in $\mathbb{R}$. Therefore $r \in f(X)$.   ∎
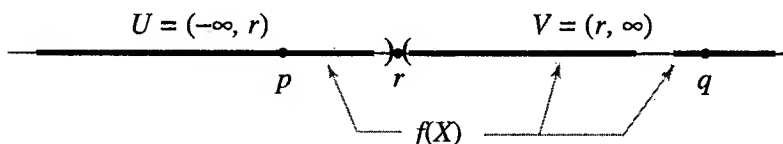
FIGURE 6.18: If $r \notin f(X)$, then $U = (-\infty, r)$ and $V = (r, \infty)$ form a separation of $f(X)$.

The Intermediate Value Theorem on $[a, b]$ follows from the general version if we let X be the interval $[a, b]$ and let p and q be $f(a)$ and $f(b)$, respectively.

EXAMPLE 6.20. Let the surface of Lake Louise be represented by a topological space X, and let $f : X \to \mathbb{R}$ be the depth of the lake beneath each point on the surface. The maximum depth of Lake Louise is 70 meters. Is there a point where the depth is 30 meters?

Note that f is a continuous function. Furthermore, X is topologically a disk, and therefore is connected. (See Exercise 6.17.) Hence the Intermediate Value Theorem applies. There are points where the value of f is 0, namely

points on the shore of the lake. There is also at least one point where the value of f is 70. Hence, given any depth between 0 and 70, there must be at least one point where the lake has that depth. In particular, there is at least one point where the depth is 30 meters.

A useful consequence of the Intermediate Value Theorem is the following one-dimensional version of the Brouwer Fixed Point Theorem:

THEOREM 6.25. *The One-Dimensional Brouwer Fixed Point Theorem.* Let $f : [-1, 1] \to [-1, 1]$ *be continuous. There exists at least one* $c \in [-1, 1]$ *such that* $f(c) = c$.

This theorem asserts that every continuous function $f : [-1, 1] \to [-1, 1]$ maps some $c \in [-1, 1]$ to itself. Such a point c is called a **fixed point** of f.

We can easily visualize why the theorem holds. (See Figure 6.19.) As we trace out the graph of f, going from the left side of the square $[-1, 1] \times [-1, 1]$ to the right side, there must be a point (c, c) where the graph intersects the line $y = x$. Such an intersection corresponds to a value c satisfying $f(c) = c$, as desired.
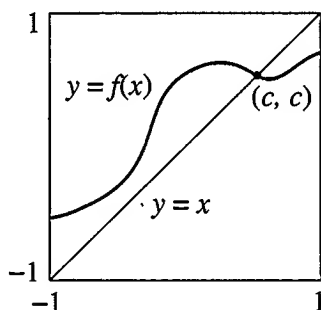
FIGURE 6.19: A fixed point of f occurs where the graph of f intersects the line $y = x$.

We make this pictorial argument formal in the following proof:

Proof. Let $f : [-1, 1] \to [-1, 1]$ be continuous. Define a function $g : [-1, 1] \to \mathbb{R}$ by $g(x) = f(x) - x$. The function g is continuous. Note that $f(-1) \geq -1$, and therefore $g(-1) \geq 0$. Similarly $g(1) \leq 0$. The Intermediate Value Theorem implies that there exists a value $c \in [-1, 1]$ such that $g(c) = 0$. For such c it follows that $f(c) = c$. Therefore there exists at least one c in $[-1, 1]$ such that $f(c) = c$, as we wished to show. ∎

The n-Dimensional Brouwer Fixed Point Theorem states that every continuous function $f : B^n \to B^n$, mapping the n-ball to itself, has a fixed point (that is, a point x where $f(x) = x$). We prove the two-dimensional version of it in Chapter 10 when we consider fixed point theory in general.

EXAMPLE 6.21. Application to Population Modeling. In the field of ecology, population models of various types—including anywhere from one to many species—are investigated both experimentally and mathematically. How a species grows and declines, how it interacts with its environment, and how it can be managed are all factors that are incorporated into and investigated via models. Many mathematical tools are used in developing and analyzing such models. When a particular model calls for a qualitative analysis, it is often topological tools that are employed.

Here we consider a straightforward single-species model exhibiting a particular type of behavior that can be identified via the Intermediate Value Theorem.

Let $y(t)$ represent the size of the species population as a function of time t. We assume that the population grows at a rate continuously dependant on the size of the present population. Thus $\frac{dy}{dt} = f(y)$ and f is continuous. The function f is the growth function. It reflects how the rate of growth depends on the population.

A simple example of such a model is given by $\frac{dy}{dt} = ky$, where $k > 0$. Here the rate of growth is proportional to the size of the population. This model applies when a population has no restriction on its resources and no predators. The family of solutions to this differential equation is given by $y = y_0 e^{kt}$. This result yields exponential growth for the population.

For other instances of the growth function $f(y)$, we may not be able to determine the general form of the solution. Approaches other than directly solving the equation are then needed to understand the behavior of the solutions.

In investigating such models, it is often of interest to find solutions that are constant in time—also known as **steady-state solutions**. Such solutions can represent the expected long-term behavior of a system, usually appearing as the limiting behavior of time-varying solutions. Thus we are interested in steady-state solutions because they often correspond to stable states toward which other solutions tend. A steady-state solution $y(t) = y^*$ occurs whenever $\frac{dy}{dt} = 0$ and thus at a value y^* such that $f(y^*) = 0$.

Let us consider a specific example. Let $y(t)$ represent the wolf population in an area of Montana where wolves are being reintroduced into the wild. We would like to know if it is possible for there to be a steady-state population of wolves.

We make two straightforward and natural assumptions about the corresponding growth function f:

(i) For small numbers of wolves, enough resources are available to support the growth of the population.

(ii) For large numbers of wolves, the resource base is not sufficient to sustain growth, and consequently the population declines.

Thus we assume that $f(y) > 0$ for small values of y, and that $f(y) < 0$ for large values of y. Let us take two such values: S, close to 0, for which $f(S) > 0$, and L, large, for which $f(L) < 0$.

Now, consider the y versus t coordinate system sketched in Figure 6.20. Solutions to the differential equation $\frac{dy}{dt} = f(y)$ have slopes specified by $f(y)$. That is, a solution curve passing through the point (t, y) has slope $f(y)$. In particular, by our assumptions, the slope of each solution curve is positive at $y = S$ and negative at $y = L$, as indicated in Figure 6.20.
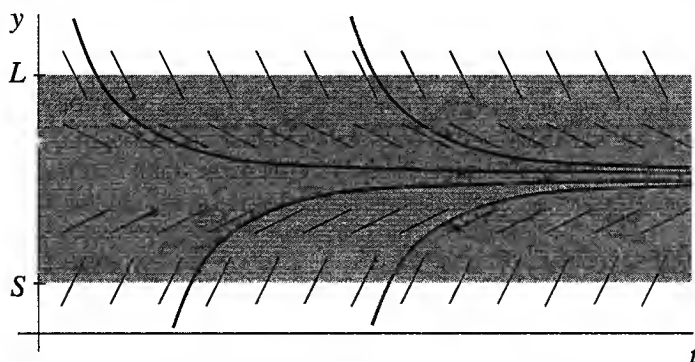
FIGURE 6.20: Slopes and solution curves described by $\frac{dy}{dt} = f(y)$ where $f(S) > 0$ and $f(L) < 0$.

In the ty-plane, let B be the band defined by

$$B = \{(t, y) \mid t \geq 0, \ S \leq y \leq L\}.$$

It is apparent in Figure 6.20 that solution curves that enter B must remain in B. What happens within the band B? Without more information about the behavior of the growth function f, not much can be concluded. However, the Intermediate Value Theorem does guarantee that, since $f(S) > 0$ and $f(L) < 0$, there exists a value $y^* \in [S, L]$ such that $f(y^*) = 0$. (See Figure 6.21.) Thus there is a steady state solution $y(t) = y^*$. Its graph is a horizontal line in the band B. Hence, under the assumptions of growth for small population values and decline for large population values, we can conclude that there is at least one population value at which the wolves survive in a steady state.
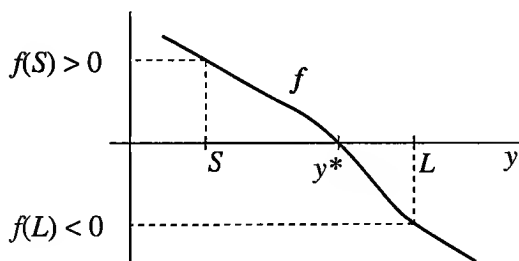
FIGURE 6.21: The growth function $f(y)$.

In the preceding example, we performed a simple qualitative analysis of a differential equation. In general, the study of the qualitative behavior of solutions of differential equations is a part of the field of dynamical systems. (We investigate another aspect of dynamical systems theory in Chapter 8.) Our previous analysis can be generalized to a straightforward dynamical systems result: Suppose $\frac{dy}{dt} = f(y)$ is defined on $[a, b]$; if f is continuous and such that $f(b) < 0$ and $f(a) > 0$, then there is a steady-state solution $y(t) = y^*$ with $y^* \in [a, b]$. This is typical of the type of qualitative conclusions about differential equation solutions that can be made using the tools of topology. Even if we are unable to specify the solutions to a differential equation, a topological analysis often allows us to make conclusions about their behavior.

The general version of the Intermediate Value Theorem has a variety of other interesting implications. For instance, we have the following theorem:

THEOREM 6.26. *Let $f : S^2 \to \mathbb{R}$ be continuous. There exists $c \in S^2$ such that $f(c) = f(-c)$.*

> ***Proof.*** Let $f : S^2 \to \mathbb{R}$ be continuous. Define $g : S^2 \to \mathbb{R}$ by $g(x) = f(x) - f(-x)$. The value of g at a point $x \in S^2$ is the difference between f at x and f at the antipode of x. Note that g is continuous, the domain S^2 is connected, and $g(x) = -g(-x)$ for all $x \in S^2$. Pick a point $p \in S^2$. If $g(p) = 0$, then $f(p) = f(-p)$ and we are done. Otherwise, $g(p)$ and $g(-p)$ are of opposite sign, and the Intermediate Value Theorem implies that there exists $q \in S^2$ such that $g(q) = 0$. For such q, it then follows that $f(q) = f(-q)$. In either case ($g(p) = 0$ or $g(p) \neq 0$), it follows that there exists $c \in S^2$ such that $f(c) = f(-c)$, establishing the desired result. ∎

EXAMPLE 6.22. Theorem 6.26 has a curious application to meteorology. Assuming that the surface of the Earth is a sphere, and surface temperature is a continuous function defined on the sphere, at any given time there is a point on Earth that has the same temperature as the point directly opposite it. Of course, we could replace temperature with any other continuous variable defined on the Earth, such as barometric pressure, relative humidity, or elevation, to obtain a corresponding result.

Theorem 6.26 can be generalized to a result for continuous functions $f : S^n \to \mathbb{R}^m$ with $n \geq m$. This general result is known as the Borsuk–Ulam Theorem. Theorem 6.26 addresses the case for $n = 2$ and $m = 1$. For $m = 1$ and an arbitrary n, the proof of Theorem 6.26 carries over exactly as presented. However, the proof of the general Borsuk–Ulam Theorem requires topological tools beyond the scope of this text.

We prove the Borsuk–Ulam Theorem for $m = 2$ and $n = 2$ in a set of supplementary exercises in Chapter 9. The $m = 2$ and $n = 2$ case applies to functions $f : S^2 \to \mathbb{R}^2$ and provides a stronger meteorology result than the one presented in Example 6.22. Specifically, at any given time there are antipodal points on the Earth's surface that have both the same temperature and the same barometric pressure!

Exercises for Section 6.3

6.32. Let $T : S^2 \to \mathbb{R}$ be defined by equating the sphere with the surface of the Earth and letting $T(x)$ be the temperature at point x on the surface at some given time. Assume that T is a continuous function. Show that if $T(\text{Anchorage}) = -30°$ and $T(\text{Honolulu}) = 80°$, there is some point on the Earth where the temperature is $0°$. Does the conclusion necessarily hold if we restrict the domain to the fifty United States? Does the conclusion hold in the United States if $T(\text{Duluth}) = -30°$ and $T(\text{Fort Lauderdale}) = 80°$?

6.33. Let $p(x)$ be an odd-degree polynomial function. Prove that $p(x) = 0$ has at least one real solution.

6.34. Determine whether or not the general version of the Intermediate Value Theorem holds when the range $\mathbb{R}$ is given each of the following topologies:
   (a) The trivial topology
   (b) The discrete topology
   (c) The lower limit topology

6.35. State and prove an intermediate value theorem for functions mapping into the digital line.

6.36. Suppose that at a given time we measure the intensity of sunlight at each point on the Earth's surface. According to Theorem 6.26, there must be a pair of points opposite each other on the Earth's surface at which the intensity of sunlight is the same. However, if it is daytime at one point, it must be nighttime at the point opposite it! Resolve the paradox.

6.37. Show that on every great circle on the surface of the Earth (obtained by intersecting the surface of the Earth with a plane through its center), there are two opposite points with the same temperature.

6.38. **(a)** Let X be connected, and assume a homeomorphism $A : X \to X$ exists such that $A \circ A(x) = x$ for all $x \in X$. Prove that, for every continuous function $f : X \to \mathbb{R}$, there exists $x \in X$ such that $f(x) = f(A(x))$.
   (b) Use the result from part (a) to prove that somewhere on a glazed doughnut there is a point that has the same thickness of glazing as the point obtained by 180-degree rotation through the central axis of the doughnut.

6.4   Path Connectedness

In this section we introduce path connectedness, the second of the two approaches to connectedness that we mentioned at the beginning of the chapter. We show that path connectedness implies connectedness. Furthermore, we exhibit a topological space (the topologist's whirlpool) that is connected but not path connected. Thus path connectedness is a stronger condition on a topological space than connectedness.

Let X be a topological space and x and y be points in X. In Section 4.3 we defined a path from x to y in X to be a continuous function $f : [0, 1] \to X$ such that $f(0) = x$ and $f(1) = y$.

DEFINITION 6.27. *A topological space X is* **path connected** *if for every* $x, y \in X$ *there is a path in X from x to y. A subset A of a topological space X is* **path connected in** *X if A is path connected in the subspace topology that A inherits from X.*

Clearly $\mathbb{R}^n$ is path connected, as is every open ball and every closed ball in $\mathbb{R}^n$.

THEOREM 6.28. *If X is a path connected space, then it is connected.*

Proof. Let X be a path connected space. We prove that X is connected by showing that it has only one component, or equivalently that every pair of points $x, y \in X$ is contained in some connected subset of X. Thus, let x and y be arbitrary points in X. Since X is path connected, there is a path in X from x to y. The image of such a path is a connected subset of X containing both x and y. Therefore every pair of points in X is contained in a connected subset of X, and it follows that X is connected. ∎

Since path connectedness implies connectedness, and since $\mathbb{R}$ in the standard topology is clearly path connected, it might appear that we now have a simpler way to prove the connectedness of $\mathbb{R}$ than that presented in Theorem 6.17. However, there is a flaw in this argument. The proof of Theorem 6.28 uses the fact that the interval $[0, 1]$ is connected, and the proof that $[0, 1]$ is connected uses the fact that $\mathbb{R}$ is connected. Thus the proof of Theorem 6.28 is actually built on the connectedness of $\mathbb{R}$. Consequently, Theorem 6.28 cannot be used to prove that $\mathbb{R}$ is connected.

On first inspection, it may be natural to think that the properties of connectedness and path connectedness are equivalent, but in fact they are not. In this next example we see a topological space that is connected but not path connected.

EXAMPLE 6.23. The topologist's whirlpool. Let A be the planar curve, expressed in (r, θ) polar coordinates as $A = \{(\frac{\theta}{\theta+1}, \theta) \mid \theta \in (0, \infty)\}$. The curve A spirals outward toward the circle S^1 as shown in Figure 6.22. Define W, the topologist's whirlpool, to be the set $W = A \cup S^1$. It is straightforward to show that W is the closure of A in $\mathbb{R}^2$.

Note that A is the continuous image of the connected space $(0, \infty)$ and therefore A is connected. Since W is the closure of A in $\mathbb{R}^2$, it too is connected.

However, W is not path connected. The proof is somewhat intricate, but well worth the effort. To prove that W is not path connected, we show that every path in W that begins in S^1 must stay in S^1, and therefore there is no path in W from a point in S^1 to a point in A. Thus, let $p : [0, 1] \to W$ be a path such that $p(0) \in S^1$. We show that $p([0, 1]) \subset S^1$, or, equivalently, that $p^{-1}(S^1) = [0, 1]$. First, note that $p^{-1}(S^1)$ is a nonempty subset of $[0, 1]$ since $p(0) \in S^1$. We claim that $p^{-1}(S^1)$ is both open and closed in $[0, 1]$. Since $[0, 1]$ is connected, it would then follow from Theorem 6.2 that $p^{-1}(S^1) = [0, 1]$. Thus we need to prove that $p^{-1}(S^1)$ is both open and closed in $[0, 1]$.
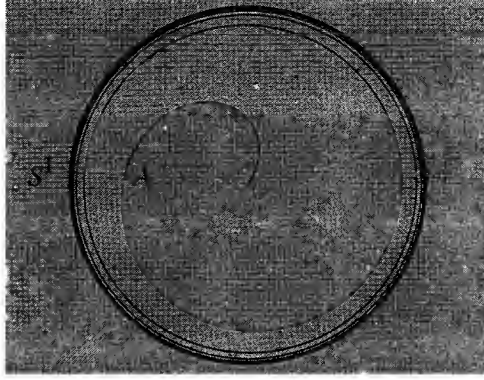
FIGURE 6.22: The topologist's whirlpool.

First, note that S^1 is closed in W and $p : [0, 1] \to W$ is continuous. Therefore $p^{-1}(S^1)$ is a closed subset of $[0, 1]$.

Let $\mathcal{B}$ be the basis for the topology on $[0, 1]$ obtained by taking the open intervals in $\mathbb{R}$ and intersecting them with $[0, 1]$. Each basis element in $\mathcal{B}$ is an interval subset of $[0, 1]$.

To see that $p^{-1}(S^1)$ is an open subset of $[0, 1]$, choose an arbitrary point $y \in p^{-1}(S^1)$. Thus $p(y) \in S^1$, and $p(y)$ has a polar-coordinate representation of the form $(1, \gamma)$. Let $N \subset \mathbb{R}^2$ be the polar neighborhood of $p(y)$ given by

$$N = \{(r, \theta) | 0.9 < r < 1.1, \gamma - 0.1 < \theta < \gamma + 0.1\}.$$

(The specific numerical values in this choice of N are not significant; we just need a small neighborhood of $p(y)$.) The set $N' = W \cap N$ is an open set in W. Thus $p^{-1}(N')$ is open in $[0, 1]$ and contains y. Therefore there is a basis element $J \in \mathcal{B}$ such that $y \in J \subset p^{-1}(N')$. Being an interval, the basis element J is connected. Hence $p(J)$ is a connected subset of N'. We show that $p(J) \subset S^1 \cap N'$. Notice that $p(J)$ has at least one point in common with $S^1 \cap N'$, namely $p(y)$.

Now, in the interval $(0.9, 1)$ we can choose a sequence (r_n) that converges to 1 and is such that the sets

$$U_n = \{(r, \theta) \in N \mid r > r_n\} \text{ and } V_n = \{(r, \theta) \in N \mid r < r_n\}$$

form a separation of N' in the plane. (See Figure 6.23.) Lemma 6.7 implies that $p(J) \subset U_n$ for all n. Thus $p(J) \subset S^1 \cap N'$. Therefore we have $y \in J \subset p^{-1}(S^1)$, and J is a basis element for the topology on $[0, 1]$. This implies that $p^{-1}(S^1)$ is open in $[0, 1]$, as we wished to show.

Thus W is not a path connected space. As we have already indicated, it is connected. Therefore the topologist's whirlpool is an example of a topological space that is connected but not path connected.
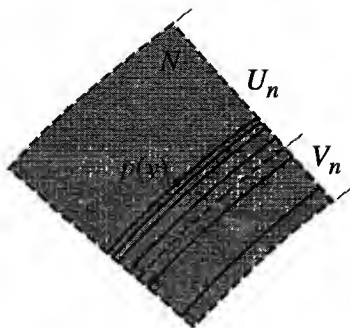
FIGURE 6.23: The sets U_n and V_n form a separation of $N' = N \cap W$.

EXAMPLE 6.24. The topologist's sine curve is the subspace of $\mathbb{R}^2$ defined by

$$T = \{(x, \sin(\tfrac{1}{x})) \mid 0 < x \le 1\} \cup \{(0, y) \mid -1 \le y \le 1\}.$$

It is shown in Figure 6.24. The topologist's sine curve is another space that is connected but not path connected. (See Exercise 6.42.)
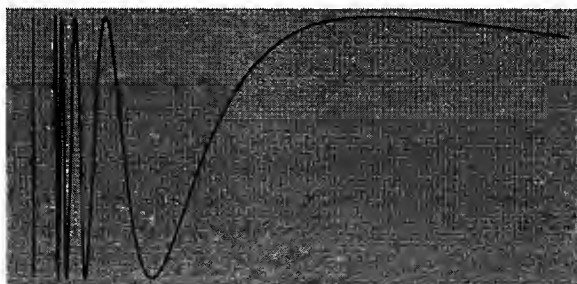
FIGURE 6.24: The topologist's sine curve.

It is not difficult to see that if we poke a hole in $\mathbb{R}^2$, the resulting space is path connected. Specifically, let p be a point in $\mathbb{R}^2$ and consider $\mathbb{R}^2 - \{p\}$. If a and b are points in $\mathbb{R}^2 - \{p\}$, and p does not lie on the line segment between a and b, then there is a straight-line path from a to b in $\mathbb{R}^2 - \{p\}$. It is given by $f(t) = (1 - t)a + tb$. Otherwise, we can find a path between a and b in $\mathbb{R}^2 - \{p\}$ by choosing a third point c that is not on the line through a, b, and p, and following a two-segment dogleg, from a to c to b, around the missing point p. (See Figure 6.25.) Such a path is given by

$$g(t) = \begin{cases} (1 - 2t)a + 2tc & \text{for } 0 \le t \le 1/2, \\ (2 - 2t)c + (2t - 1)b & \text{for } 1/2 \le t \le 1. \end{cases}$$
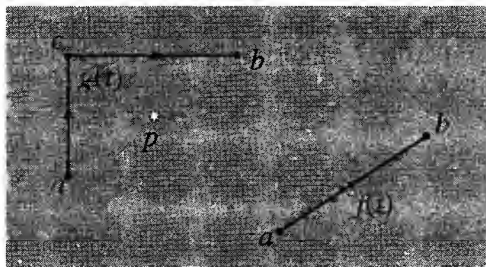
FIGURE 6.25: The space $\mathbb{R}^2 - \{p\}$ is path connected.

Note that g is continuous by the Pasting Lemma. Thus $\mathbb{R}^2 - \{p\}$ is path connected.

The preceding argument can also be used to show that for every $n \geq 2$ and $p \in \mathbb{R}^n$, the space $\mathbb{R}^n - \{p\}$ is path connected. Moreover, we need not limit ourselves to one pinprick in $\mathbb{R}^n$. Even putting a countably infinite number of pinpricks in $\mathbb{R}^n$ results in a space that is path connected. (See Exercises 6.43 and 6.44.)

The following theorem indicates that the image of a path connected space under a continuous function is path connected. This implies that path connectedness is preserved under homeomorphism and therefore is a topological property.

THEOREM 6.29. *Assume that $f : X \to Y$ is continuous and X is path connected. Then $f(X)$ is a path connected subspace of Y.*

Proof. Let p and q be points in $f(X)$. Pick points $x \in f^{-1}(\{p\})$ and $y \in f^{-1}(\{q\})$. Since X is path connected, there exists a path $g : [0, 1] \to X$ from x to y. Then $f \circ g$ is a path in $f(X)$ from p to q. ∎

Now, as is the case with connectedness, a topological space X can be partitioned into maximal path connected subsets. These are known as the path components of X. We make this precise in what follows.

Given a topological space X, define a relation on X by $x \sim_p y$ if there exists a path in X from x to y. This relation is an equivalence relation, which we now check.

(i) Is it true that $x \sim_p x$? Yes, we can take the path $f : [0, 1] \to X$ given by $f(t) = x$ for all t.

(ii) Is it true that $x \sim_p y$ implies $y \sim_p x$? Yes. If $f : [0, 1] \to X$ is a path from x to y, then $g : [0, 1] \to X$, defined by $g(t) = f(1 - t)$, is a path from y to x.

(iii) Suppose $x \sim_p y$ and $y \sim_p z$. Is it true that $x \sim_p z$? Yes. We know that there are paths f and g between x and y and between y and z, respectively. Form a new path $h : [0, 1] \to X$, defined by

$$h(t) = \begin{cases} f(2t) & \text{for } 0 \leq t \leq 1/2, \\ g(2t - 1) & \text{for } 1/2 \leq t \leq 1. \end{cases}$$

Then $h(0) = x$ and $h(1) = z$. Since $f(1/2) = y = g(1/2)$, the Pasting Lemma implies that h is continuous. Thus, h is a path from x to z.

Therefore $\sim_p$ is an equivalence relation. As usual, the equivalence classes form a partition of X.

DEFINITION 6.30. *The equivalence classes under the equivalence relation $\sim_p$ are called the **path components** of X.*

The path components of X are path connected, and every path connected subset of X is a subset of a path component of X. (See Exercise 6.47.) In this sense, the path components of a topological space X are the maximal path connected subsets of X.

EXAMPLE 6.25. Let X_n be the subset of the plane consisting of the spiral $\{(\frac{\theta}{\theta+1}, \theta) \in \mathbb{R}^2 \mid \theta \in (0, \infty)\}$ and n equally-spaced points in the circle S^1. We illustrate X_6 in Figure 6.26. With arguments similar to those in the topologist's whirlpool example, we can show that X_n is connected but not path connected. It follows that X_n has one component, and it is straightforward to prove that X_n has $n + 1$ path components.
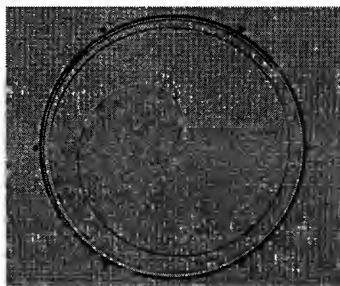
FIGURE 6.26: The set X_6 consists of a spiral and six equally-spaced points in the circle surrounding it.

It is generally true that the number of components of a topological space does not exceed the number of path components since each path component must be a subset of some component of the space. (See Exercise 6.49.)

Exercises for Section 6.4

6.39. Let $X = \{a, b\}$ and assume that X has the topology $\mathcal{T} = \{\varnothing, \{a\}, X\}$. Show that X is path connected in this topology.

6.40. Prove that the digital line is path connected.

6.41. A set $C \subset \mathbb{R}^n$ is said to be **star convex** if there exists a point $p^* \in C$ such that for every $p \in C$, the line segment in $\mathbb{R}^n$ joining p^* and p lies in C. Prove that if $C \subset \mathbb{R}^n$ is star convex, then C is path connected in $\mathbb{R}^n$.

6.42. Prove that the topologist's sine curve is connected but not path connected.

6.43. Let $A = \mathbb{R}^2 - C$ where C is a countable subset of $\mathbb{R}^2$. Prove that A is path connected. (Hint: Given a point $\alpha \in A$, there are many lines through α that miss C entirely. Why?)

6.44. Let $A = \mathbb{R}^n - C$ where C is a countably infinite subset of $\mathbb{R}^n$. Prove that A is path connected. (Hint: Prove this by induction, using Exercise 6.43 as an initial step.)

6.45. Prove that the n-sphere S^n is path connected for $n \geq 1$. (Hint: Find a surjective continuous function from a path connected space to S^n.)

6.46. Prove that if $\{A_\alpha\}_{\alpha \in A}$ is a collection of path connected subsets of a topological space X, and $\bigcap_{\alpha \in A} A_\alpha$ is nonempty, then $\bigcup_{\alpha \in A} A_\alpha$ is path connected.

6.47. Let X be a topological space.
  (a) Prove that each path component of X is path connected.
  (b) Prove that each path connected subset of X is a subset of some path component of X.

6.48. Provide an example showing that the path components of a topological space X need not be open sets in X, nor closed sets in X.

6.49. Prove that each path component of a topological space X is a subset of some component of X.

6.50. Provide an example of a connected topological space having uncountably many path components.

6.51. Prove that if $f : X \to Y$ is a homeomorphism and C is a path component of X, then $f(C)$ is a path component of Y.

6.52. Let $X_1, \ldots, X_n$ be path connected. Prove that the corresponding product space $X_1 \times \ldots \times X_n$ is path connected.

6.5 *Automated Guided Vehicles*

Automated guided vehicles are mobile robots that are used to transport materials from location to location in a manufacturing facility. Part of the challenge in designing and constructing such a facility is properly setting up mobile robot routes so that the robots can maneuver in an efficient and safe manner. The tools and concepts of topology are naturally employed in this planning process. In this section we consider some simple examples.

We model each robot with a point that moves through a topological space representing the robot routes in the factory. To begin, suppose we have two robots A and B that move on a line represented by $\mathbb{R}$, as shown in Figure 6.27. Let x_A indicate the location of robot A and x_B the location of robot B. The configuration space for the two robots (see Section 3.5) is the space

$$C = \{(x_A, x_B) \mid x_A \in \mathbb{R},\ x_B \in \mathbb{R}\} = \mathbb{R} \times \mathbb{R} = \mathbb{R}^2.$$
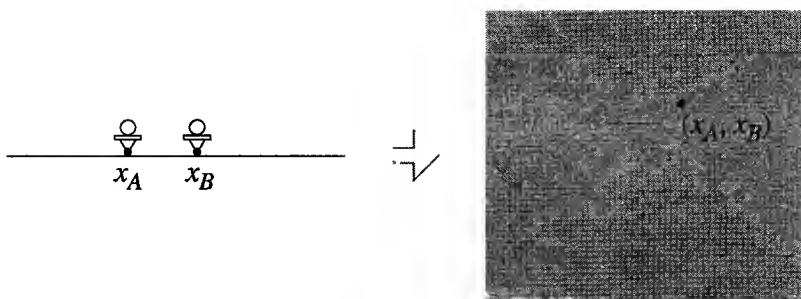
FIGURE 6.27: Two robots moving on the line yield a configuration space that is $\mathbb{R}^2$.

To prevent robot collisions, we do not allow the two robots to occupy the same point at the same time. That is, we never let $x_A = x_B$. The resulting space of permitted configurations is then

$$SC = \{(x_A, x_B) \mid x_A \in \mathbb{R}, \ x_B \in \mathbb{R}, \ x_A \neq x_B\}.$$

To distinguish this space from the configuration space C, we call this the **safe configuration space** for the robots.

The points in the plane that are excluded from the safe configuration space comprise the set $\{(x, x) \mid x \in \mathbb{R}\}$. This set is called the **diagonal** of the plane and is denoted by Δ. Note that $SC = C - \Delta$. In this case the safe configuration space is the plane with the diagonal removed. (See Figure 6.28.) It is important to note that this space is not connected and therefore not path connected. We will shortly see the significance of this fact.
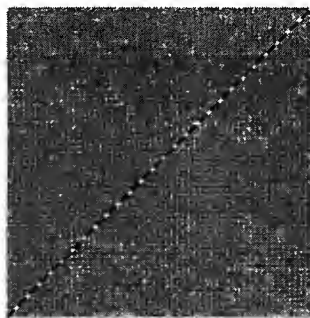
FIGURE 6.28: The safe configuration space for two robots on the line is the plane with the diagonal removed.

In the general case, suppose that we have n robots in space X. The **safe configuration space** is defined to be the space

$$SC^n(X) = (X \times X \times \ldots \times X) - \Delta,$$

where

$$\Delta = \{(x_1, x_2, \ldots, x_n) \mid x_i = x_j \text{ for some } i \neq j\}.$$

Here we are taking the product of n copies of X and removing the diagonal Δ. As in the two-robot case, by removing the set Δ we are excluding configurations where two or more robots occupy the same point and are therefore eliminating the possibility of robot collisions.

To model a relocation of the robots in space X, we use a path in $SC^n(X)$. Specifically, assume that our n robots are initially positioned so that robot j is located at I_j, and we wish to move the robots so that the final position of robot j is F_j. We define a **relocation** of the n robots from the initial configuration $I = (I_1, \ldots, I_n)$ to the final configuration $F = (F_1, \ldots, F_n)$ to be a path $p : [0, 1] \to SC^n(X)$ such that $p(0) = I$ and $p(1) = F$. The existence of such a path ensures that the robots can be moved from the initial configuration to the final configuration without collision.

Given two configurations of the robots, $M = (M_1, \ldots, M_n)$ and $N = (N_1, \ldots, N_n)$, we say that N is **attainable** from M if there is a relocation of the robots from M to N. If the safe configuration space is path connected, then every configuration is attainable from every other one, and we say that the robots are **freely transportable in** X.

EXAMPLE 6.26. Returning to the example of two robots on the line, the safe configuration space is $SC^2(\mathbb{R}) = \mathbb{R}^2 - \Delta$. As we already observed, this space is not path connected, so the two robots are not freely transportable in $\mathbb{R}$. This is intuitively clear, since no configuration with robot A to the left of robot B is attainable from any of the configurations with robot B to the left of robot A.

This intuitive observation is reflected in the structure of the safe configuration space. In particular, the path components of $SC^2(\mathbb{R})$ are the sets

$$U = \{(x_A, x_B) \mid x_A < x_B \in R\} \text{ and } V = \{(x_A, x_B) \mid x_A > x_B \in R\},$$

which lie above and below the diagonal, respectively, in Figure 6.28. The set U represents all of the configurations where robot A lies to the left of robot B, and the set V represents all of the configurations where robot B lies to the left of robot A. Since there is no path in $SC^2(\mathbb{R})$ from U to V, it follows—as we have observed intuitively—that no configuration with robot A to the left of robot B is attainable from any of the configurations with robot B to the left of robot A. Nonetheless, since there is a path in U between every pair of configurations in U, it follows that each configuration with robot A to the left of robot B is attainable from every other such configuration via a relocation that keeps robot A to the left of robot B. The same holds true for each pair of configurations in V.

In Figure 6.29 we illustrate two relocations of the two robots from an initial configuration (I_A, I_B) to a final configuration (F_A, F_B) in U. Since the configurations are in U, robot B is to the right of robot A. In the first relocation, both robots move to the right, with robot B traveling further overall than robot A, a situation that is reflected in the fact that the slope of the segment between (I_A, I_B) and (F_A, F_B) is greater than 1. In the second relocation, the robots move toward each other, and since they are moving in opposite directions, the slope of the segment between (I_A, I_B) and (F_A, F_B) is negative.
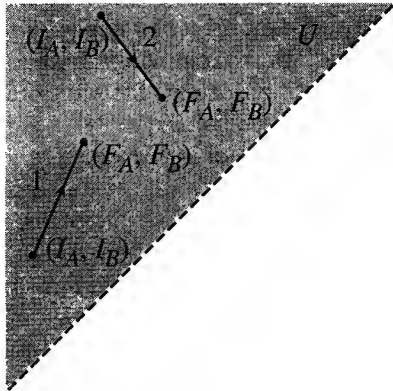
FIGURE 6.29: Relocations in U.

EXAMPLE 6.27. Consider two robots on the circle S^1. Here the safe configuration space $SC^2(S^1)$ is a torus, $S^1 \times S^1$, with the diagonal removed. To picture the safe configuration space, consider Figure 6.30. There we represent the torus as a square with opposite edges identified. The removed diagonal is represented by the dashed line. Removal of the diagonal results in a space formed by gluing two triangles along the edges as shown. The corresponding space is homeomorphic to $(0, 1) \times S^1$, an open annulus.
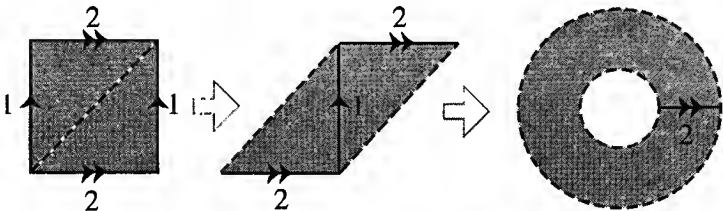
FIGURE 6.30: The safe configuration space $SC^2(S^1)$ is an open annulus.

An open annulus is path connected since it is homeomorphic to a product of two path connected spaces ($(0, 1)$ and S^1) and since the product of path connected spaces is path connected. (See Exercise 6.52.) Therefore we can conclude that two robots on S^1 are freely transportable.

EXAMPLE 6.28. This is based on an example from [Abr2]. Consider the configuration space for two robots A and B on the Y-shaped topological graph shown in Figure 6.31. We denote the space by Y and the three edges of Y by α, β, and γ. The three edges intersect at the vertex v.

The safe configuration space is $SC^2(Y) = (Y \times Y) - \Delta$. In order to picture $SC^2(Y)$ we break it up into three subsets: $(Y \times \alpha) - \Delta$, $(Y \times \beta) - \Delta$, and $(Y \times \gamma) - \Delta$, then we examine how the three subsets glue together to form $SC^2(Y)$.
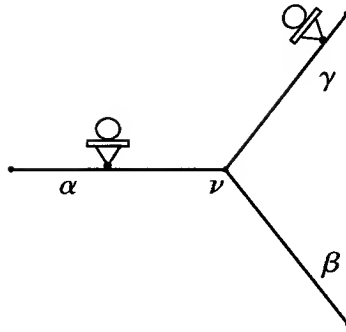
FIGURE 6.31: Two robots on the space Y.

In Figure 6.32 we depict $(Y \times \alpha) - \Delta$. The set of configurations corresponding to robot B fixed at v and robot A located anywhere else on Y is represented by the edges at the bottom of the pictured space. Along the edge marked with a single arrow, robot B is fixed at v and robot A lies along edge α. We can interpret the edges marked with a double or triple arrow in a similar fashion. Moving vertically in the illustration of $(Y \times \alpha) - \Delta$ corresponds to moving robot B outward along edge α and keeping robot A fixed. Because we exclude the diagonal line segment that corresponds to the robots coinciding on edge α, there are two components to $(Y \times \alpha) - \Delta$: the triangle below the diagonal line segment, and the rest of $(Y \times \alpha) - \Delta$, as shown in the figure.
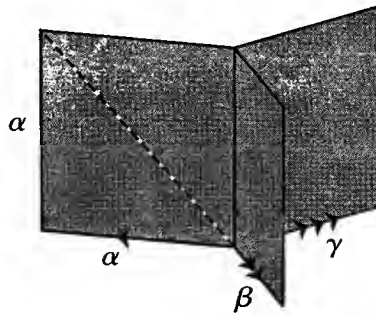
FIGURE 6.32: The subset $(Y \times \alpha) - \Delta$ of the safe configuration space $(Y \times Y) - \Delta$.

In Figure 6.33 we show $(Y \times \beta) - \Delta$ and $(Y \times \gamma) - \Delta$. The situation there is similar to that shown in Figure 6.32. It is important to realize that the edges marked with a single arrow in $(Y \times \beta) - \Delta$ and $(Y \times \gamma) - \Delta$ correspond to the edge marked with a single arrow in $(Y \times \alpha) - \Delta$. Each of these edges corresponds to the configurations where robot B is fixed at v and robot A is somewhere on edge α. Therefore these edges are glued together in the construction of $(Y \times Y) - \Delta$ from $(Y \times \alpha) - \Delta$, $(Y \times \beta) - \Delta$, and $(Y \times \gamma) - \Delta$. In a similar manner, the edges marked with double arrows are glued together in the construction of $(Y \times Y) - \Delta$, as are the edges marked with triple arrows.
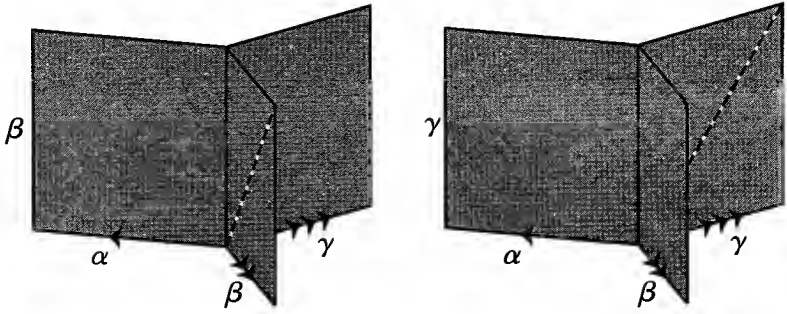
FIGURE 6.33: The spaces $(Y \times \beta) - \Delta$ and $(Y \times \gamma) - \Delta$.

Taking $(Y \times \alpha) - \Delta$, $(Y \times \beta) - \Delta$, and $(Y \times \gamma) - \Delta$ and gluing them together so that the arrows appropriately match, we obtain the safe configuration space $SC^2(Y) = (Y \times Y) - \Delta$ shown in Figure 6.34. The parts labeled 1, 2, and 3, come from $(Y \times \alpha) - \Delta$, $(Y \times \beta) - \Delta$, and $(Y \times \gamma) - \Delta$, respectively. The hole in the middle is the deleted diagonal point corresponding to having both robots located at v. This safe configuration space $SC^2(Y)$ is path connected. Therefore a pair of robots is freely transportable on Y.
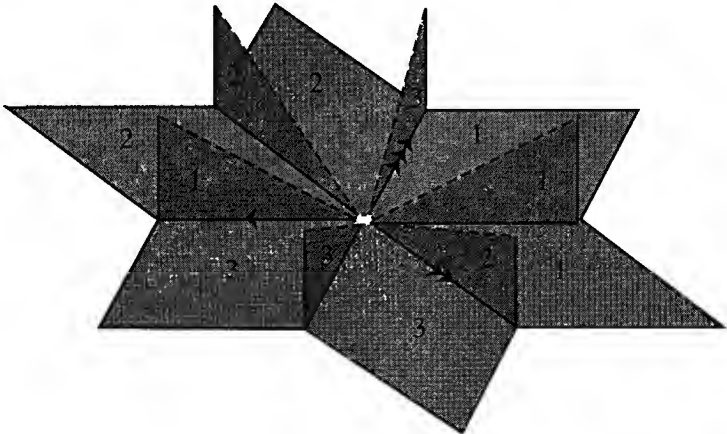
FIGURE 6.34: The safe configuration space $SC^2(Y) = (Y \times Y) - \Delta$.

In Figure 6.35, looking directly down on $SC^2(Y)$ as shown in Figure 6.34, we depict a relocation of the robots that switches their positions, starting with robot A near the end of α and robot B near the end of β. The path corresponds to the following maneuver of the robots:

(i) Keeping robot B fixed, move robot A to point v and then onto edge γ.

(ii) Keeping robot A fixed, move robot B to point v and then out to its final position on edge α.

(iii) Keeping robot B fixed, move robot A back to point v and then out to its final position on edge β.
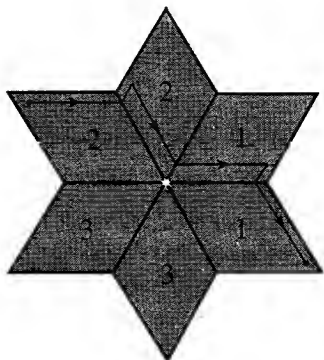
FIGURE 6.35: A relocation that switches the positions of two robots on Y.

Exercises for Section 6.5

6.53. Consider $SC^2(\mathbb{R})$, the safe configuration space for two robots on a line. For each of the following cases, sketch a line segment in $SC^2(\mathbb{R})$ illustrating the described relocation:

  (a) Robot B is to the left of robot A and they move in opposite directions, toward each other.

  (b) Robot B is to the left of robot A and they both move to the left, with robot B traveling further than robot A.

  (c) Robot A is to the left of robot B and they move in opposite directions, away from each other.

  (d) Robot A is to the left of robot B and they both move to the right, with robot A traveling further than robot B.

6.54. Describe a general procedure for performing a relocation from any given configuration of two robots on the circle to any other configuration.

6.55. Describe the safe configuration space for three robots moving on $\mathbb{R}$, indicating how many different path components make up the space and discussing which relative positions of the robots correspond to each path component.

6.56. Suppose that we have two robots on the line, and there is a particular coordinate x_0 on the line that they cannot occupy. Sketch and describe the safe configuration space for this situation. Indicate how many different path components make up the space and discuss which relative positions of the robots correspond to each path component.

6.57. Sketch a path on $SC^2(Y)$ representing a relocation that switches the robots' positions, starting with both robots located on edge γ. Describe the corresponding robot maneuvers.

6.58. Consider $SC^2(P)$, the safe configuration space for two robots on the space P shown on the left in Figure 6.36.

  (a) The space $SC^2(P)$ can be obtained as a quotient space of $SC^2(Y)$ by gluing together parts of $SC^2(Y)$. On an illustration of $SC^2(Y)$, indicate which parts are glued together to obtain $SC^2(P)$. (Hint: Think about how to obtain P from Y by identifying points on Y.)

    (b) On a representation of $SC^2(P)$, sketch a path illustrating a relocation that switches the robots' positions, starting with robot A at the top of P and robot B at the bottom. Describe the corresponding robot maneuvers.

6.59. Consider $SC^2(X)$, the safe configuration space for two robots on the space X shown on the right in Figure 6.36.

    (a) Sketch a representation of $SC^2(X)$, showing how it can be constructed by gluing together subsets in a manner similar to our construction of $SC^2(Y)$.

    (b) On your representation of $SC^2(X)$, sketch a path illustrating a relocation that switches the robots' positions, starting with robot A at the top left of X and robot B at the top right. Describe the corresponding robot maneuvers.
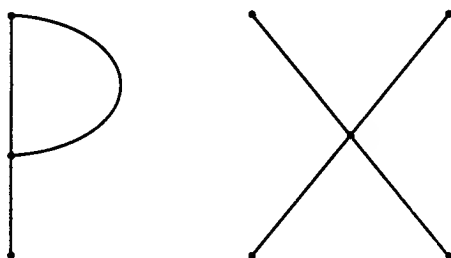
FIGURE 6.36: The spaces P and X.

Chapter 7

Compactness

So far we have introduced two of the "Three Cs" of introductory topology: continuity and connectedness. In this chapter, we present the third: compactness. This concept is not as intuitive as continuity or connectedness. In $\mathbb{R}^n$, the compact sets are the closed and bounded sets, but in a general topological space the compact sets are not as simple to describe. In fact, a formal definition for compactness in a general topological space took some time for topologists to establish. Several definitions were suggested during the development of topology in the early twentieth century. Ultimately, topologists settled on the definition of compactness proposed in 1923 by Pavel Sergeevich Alexandroff (1896–1982) and Pavel Urysohn; it is presented here in the first section.

We define compactness and examine some related results and examples in Section 7.1. Compactness in metric spaces is addressed in Section 7.2. In Section 7.3, we prove the Extreme Value Theorem and then use it to establish a number of other useful results. In Section 7.4, we investigate limit point compactness, a property that is closely related to compactness. In particular, we prove that in a metric space, limit point compactness is equivalent to compactness. Finally, in Section 7.5, we introduce the one-point compactification, a construction that enables us to view a noncompact space X as a subspace of a compact space that is obtained by adding a single point to X and defining an appropriate topology.

7.1 Open Coverings and Compact Spaces

Consider the subsets of $\mathbb{R}^2$ in the standard topology depicted in Figure 7.1. Our goal in this chapter is to establish the distinction between the subsets in the figure that are labeled "compact" and the subsets that are labeled "noncompact."
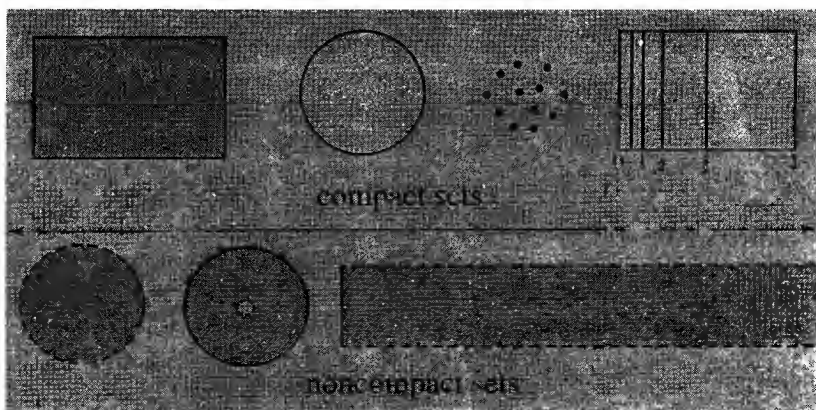
FIGURE 7.1: Compact and noncompact subsets of $\mathbb{R}^2$.

We start with some important definitions associated with compactness.

DEFINITION 7.1. *Let A be a subset of a topological space X, and let $\mathcal{O}$ be a collection of subsets of X.*

> *(i) The collection $\mathcal{O}$ is said to **cover** A or to be a **cover** of A if A is contained in the union of the sets in $\mathcal{O}$.*
>
> *(ii) If $\mathcal{O}$ covers A, and each set in $\mathcal{O}$ is open, then we call $\mathcal{O}$ an **open cover** of A. (See Figure 7.2.)*
>
> *(iii) If $\mathcal{O}$ covers A, and $\mathcal{O}'$ is a subcollection of $\mathcal{O}$ that also covers A, then $\mathcal{O}'$ is called a **subcover** of $\mathcal{O}$.*
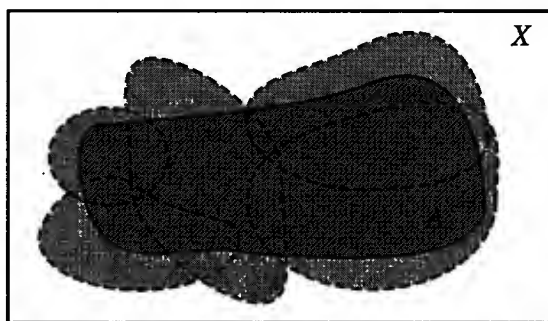
FIGURE 7.2: An open cover of A.

By the definition of a basis, it follows that every basis for a topological space X is an open cover of X.

The two collections of intervals

$$\mathcal{O}_1 = \{\ldots, (-1, 1), (0, 2), (1, 3), \ldots\} \text{ and } \mathcal{O}_2 = \{(-\infty, 1), (0, \infty)\}$$

are both open covers of $\mathbb{R}$. As we see from these collections, we can have open covers that consist of infinitely many sets, and we can have open covers that consist of finitely many sets. We are interested in spaces for which every open cover containing infinitely many sets can be reduced to a subcover containing finitely many sets. Specifically, we have the following definition:

DEFINITION 7.2. *A topological space X is **compact** if every open cover of X has a finite subcover.*

EXAMPLE 7.1. The real line $\mathbb{R}$ in the standard topology is not compact since

$$\mathcal{O} = \{\ldots, (-1, 1), (0, 2), (1, 3), \ldots\}$$

is an open cover, but no finite subcollection of $\mathcal{O}$ covers $\mathbb{R}$. (See Figure 7.3.)
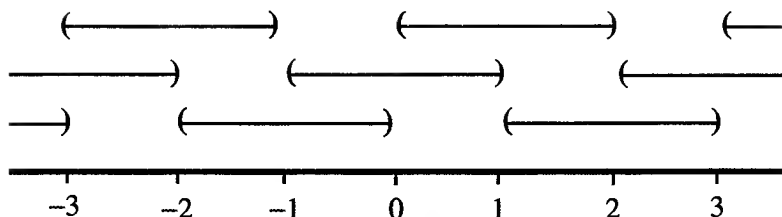
FIGURE 7.3: An open cover of $\mathbb{R}$ with no finite subcover.

EXAMPLE 7.2. Let $X = \{x_1, \ldots, x_n\}$ be a topological space that contains only finitely many points. Then X is compact since there can be only finitely many open sets in the topology on X, and therefore every open cover of X is already finite.

We extend the definition of compactness to subsets of a topological space, as follows:

DEFINITION 7.3. *Let X be a topological space, and assume $A \subset X$. Then A is said to be* **compact in** X *if A is compact in the subspace topology inherited from X.*

The following lemma allows us to check whether or not a subspace A of a topological space X is compact by considering covers of A that are made up of open sets in X rather than covers of A that are made up of open sets in the subspace topology on A:

LEMMA 7.4. *Let X be a topological space, and assume $A \subset X$. Then A is compact in X if and only if every cover of A by sets that are open in X has a finite subcover.*

> ***Proof.*** Let A be compact in X, and suppose that $\mathcal{O}$ is a cover of A by open sets in X. Then $\mathcal{O}' = \{U \cap A \mid U \in \mathcal{O}\}$ is a cover of A by open sets in A. Hence, there exists a finite subcover $\{U_1 \cap A, U_2 \cap A, \ldots, U_n \cap A\}$ of $\mathcal{O}'$. But then $\{U_1, U_2, \ldots, U_n\}$ is a finite subcover of $\mathcal{O}$. Therefore every cover of A by open sets in X has a finite subcover.
>
> Conversely, suppose every cover of A by sets that are open in X has a finite subcover. Let $\mathcal{O} = \{V_\beta\}_{\beta \in B}$ be a cover of A by open sets in A. Then, by definition of the subspace topology, for each V_β there is an open set U_β in X such that $V_\beta = U_\beta \cap A$. It follows that the collection $\mathcal{O}' = \{U_\beta\}_{\beta \in B}$ is a cover of A by open sets in X. Since $\mathcal{O}'$ has a finite subcover $\{U_{\beta_1}, \ldots, U_{\beta_n}\}$, it follows that $\{V_{\beta_1}, \ldots, V_{\beta_n}\}$ is a finite subcover of $\mathcal{O}$. Thus every cover of A by open sets in A has a finite subcover, and therefore A is compact. ∎

EXAMPLE 7.3. The subset $A = \{0\} \cup \{\frac{1}{n} \mid n \in \mathbb{Z}_+\}$ is compact in $\mathbb{R}$. To see this, let $\mathcal{O}$ be a cover of A by open sets in $\mathbb{R}$. There exists at least one open set U_0 in $\mathcal{O}$ that contains the point 0. Such an open set contains all but at most finitely many of the points in A. (See Figure 7.4.) If U_0 contains all of the points in A, then U_0, by itself, is a finite subcover of $\mathcal{O}$. Otherwise, let $1/m$ be the smallest of the points in A that are not in U_0. For each point $1/i$, there is an open set U_i in $\mathcal{O}$ that contains it. It follows that the finite collection $\{U_0, U_1, ..., U_m\}$ is a subcover of $\mathcal{O}$. Thus, A is compact in $\mathbb{R}$.
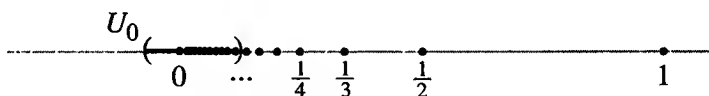

FIGURE 7.4: The set U_0 contains all but finitely many points from A.

EXAMPLE 7.4. Consider $(0, 1]$ as a subspace of $\mathbb{R}$. This space is not compact. The collection $\mathcal{O} = \{(\frac{1}{n}, 2) \mid n \in \mathbb{Z}_+\}$ is a cover of $(0, 1]$ by sets that are open in $\mathbb{R}$. There is no finite subcollection of $\mathcal{O}$ that covers $(0, 1]$, and therefore $(0, 1]$ is not compact as a subspace of $\mathbb{R}$.

Two homeomorphic spaces are either both compact or both noncompact. Every open cover $\mathcal{O}$ of one is sent to an open cover $\mathcal{O}'$ of the other by a homeomorphism. Similarly, every finite subcover of $\mathcal{O}$ is sent to a finite subcover of $\mathcal{O}'$ by the homeomorphism.

Since $\mathbb{R}$ and $(0, 1]$ are not compact, it follows that every interval of the form (a, b), $(-\infty, b)$, (a, ∞), $[a, b)$, $[a, \infty)$, $(a, b]$, or $(-\infty, b]$ is not compact as well. What about intervals in the form $[a, b]$? In the next section, we will show that these intervals are compact in the standard topology.

We actually do not need the full strength of a homeomorphism to ensure that a compact space maps to a compact space. The next theorem asserts that, as with connectedness and path connectedness, compactness is preserved by continuous functions.

THEOREM 7.5. *Let $f : X \to Y$ be continuous, and let A be compact in X. Then $f(A)$ is compact in Y.*

> **Proof.** Let $f : X \to Y$ be continuous, and assume that A is compact in X. To show that $f(A)$ is compact in Y, let $\mathcal{O}$ be a cover of $f(A)$ by open sets in Y. Then $f^{-1}(U)$ is open in X for every open set U in $\mathcal{O}$. Hence $\mathcal{O}' = \{f^{-1}(U) \mid U \in \mathcal{O}\}$ is a cover of A by open sets in X. Since A is compact, Lemma 7.4 implies that there is a finite subcollection of $\mathcal{O}'$, say $\{f^{-1}(U_1), ..., f^{-1}(U_n)\}$, that covers A. Then the collection of open sets $\{U_1, ..., U_n\}$ in $\mathcal{O}$ covers $f(A)$. Thus, $\mathcal{O}$ has a finite subcover, implying that $f(A)$ is compact in Y. ∎

One immediate consequence of Theorem 7.5 is that a quotient space of a compact space X is compact since it is the image of X under a quotient map, and a quotient map is a continuous function.

The following theorem provides some results on the compactness of unions and intersections of compact sets:

THEOREM 7.6. *Let X be a topological space.*

> *(i) If $C_1, \dots, C_n$ are each compact in X, then $\bigcup_{j=1}^n C_j$ is compact in X.*
>
> *(ii) If X is Hausdorff, and $\{C_\alpha\}_{\alpha \in A}$ is a collection of sets that are compact in X, then $\bigcap_{\alpha \in A} C_\alpha$ is compact in X.*

Proof. See Exercise 7.2.   ∎

An arbitrary union of compact sets need not be compact. (See Exercise 7.3.) Also, if we drop the assumption that X is Hausdorff in the second part of Theorem 7.6, then an intersection of compact sets need not be compact. (See Exercise 7.18.)

The next two theorems show that being closed and being compact are closely related properties.

THEOREM 7.7. *Let X be a topological space and let D be compact in X. If C is closed in X, and $C \subset D$, then C is compact in X.*

> *Proof.* Let D be compact in the topological space X. Suppose that C is closed in X and $C \subset D$. Further, suppose that $\mathcal{O}$ is a cover of C by sets that are open in X. The set $X - C$ is open. Add $X - C$ to the collection $\mathcal{O}$ to obtain a new collection $\mathcal{O}' = \mathcal{O} \cup \{X - C\}$. (See Figure 7.5.) The collection $\mathcal{O}'$ is an open cover of X and therefore is an open cover of D. Since D is compact in X, there is a finite subcollection of $\mathcal{O}'$ that covers D. The set C is covered by those sets in the finite subcover of $\mathcal{O}'$ that were originally in $\mathcal{O}$. Therefore there is a finite subcollection of $\mathcal{O}$ that covers C, implying that C is compact in X.   ∎
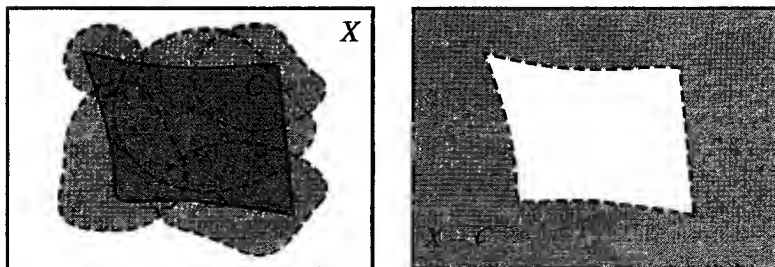

FIGURE 7.5: Adding $X - C$ to an open cover of C results in an open cover of the whole space X.

Theorem 7.7 indicates that a closed subset of a compact space is compact. The converse relationship, however, does not generally hold. A compact set in a topological space is not necessarily a closed set, as demonstrated in the next example.

EXAMPLE 7.5. Consider $\mathbb{R}_{fc}$, the real line in the finite complement topology. Every subset of $\mathbb{R}_{fc}$ is compact. (See Exercise 7.1.) Aside from the whole set itself, every other infinite set $A \subset \mathbb{R}_{fc}$ is not closed, since the complement of such a set A is not open. Therefore there are subsets of $\mathbb{R}_{fc}$ that are compact but not closed.

Although compact sets in a topological space are not necessarily closed, the following theorem indicates that there are general circumstances under which compact sets are automatically closed.

THEOREM 7.8. *Let X be a Hausdorff topological space and A be compact in X. Then A is closed in X.*

Proof. Let A be compact in the Hausdorff space X. To show that A is closed, we prove that $X - A$ is open. Thus, let $x \in X - A$ be arbitrary. We show that there is an open set U such that $x \in U \subset X - A$.

Since X is Hausdorff, we know that for each $a \in A$, there exist disjoint open sets U_a and V_a such that $x \in U_a$ and $a \in V_a$. (See Figure 7.6.) Then $\mathcal{O} = \{V_a\}_{a \in A}$ is an open cover of A. Because A is compact, there is a finite subcover $\{V_{a_1}, ..., V_{a_n}\}$ of $\mathcal{O}$. Let $V = \bigcup_{i=1}^{n} V_{a_i}$ and $U = \bigcap_{i=1}^{n} U_{a_i}$. Then U and V are open sets such that $A \subset V$ and $x \in U$. Furthermore, since U_{a_i} and V_{a_i} are disjoint for each i, it follows that U and V are disjoint as well. Thus U and A are disjoint, and therefore there exists an open set U such that $x \in U \subset X - A$, as we wished to show. Hence $X - A$ is open, implying that A is closed. ∎
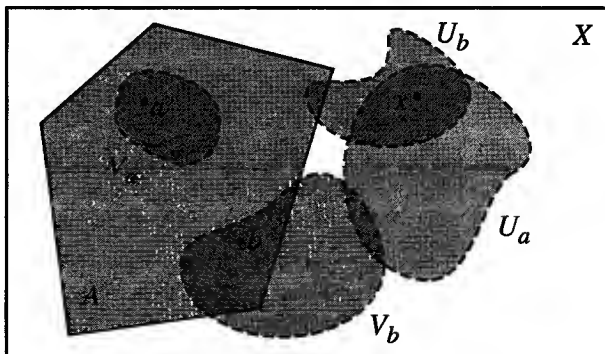

FIGURE 7.6: Using the Hausdorff property to separate x from points in A.

The next lemma will help us subsequently prove that a finite product of compact spaces is compact.

LEMMA 7.9. *The Tube Lemma.* *Let X and Y be topological spaces, and assume that Y is compact. If $x \in X$, and U is an open set in $X \times Y$ containing $\{x\} \times Y$, then there exists a neighborhood W of x in X such that $W \times Y \subset U$.*

The lemma asserts that if an open set in $X \times Y$ contains a slice $\{x\} \times Y$ of $X \times Y$, then there is an open tube $W \times Y$ containing the slice and contained in the open set. (See Figure 7.7.) Notice that, without the assumption that Y is compact, the lemma does not necessarily hold. (See Exercise 7.12.)
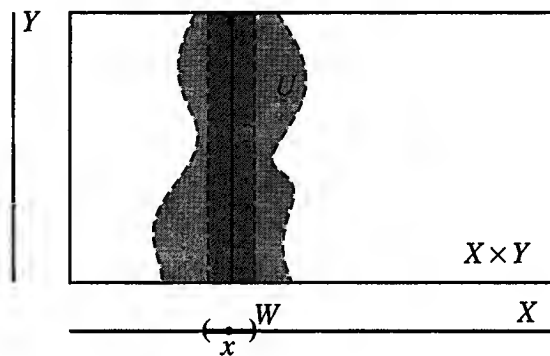
FIGURE 7.7: The tube $W \times Y$ is contained in the open set U.

Proof. For each $y \in Y$ pick open sets W_y in X and V_y in Y such that $(x, y) \in W_y \times V_y \subset U$. The collection of sets $\{V_y\}_{y \in Y}$ is an open cover of Y. Since Y is compact, finitely many of these sets cover Y, say $V_{y_1}, \dots, V_{y_n}$. Let $W = \bigcap_{i=1}^{n} W_{y_i}$. Then W is open in X, and W contains x because each W_y contains x. Note that

$$W \times Y \subset \bigcup_{i=1}^{n}(W_{y_i} \times V_{y_i}) \subset U.$$

Therefore $W \times Y$ contains $\{x\} \times Y$ and is contained in U, as desired. ∎

THEOREM 7.10. *If X and Y are compact topological spaces, then the product $X \times Y$ is compact.*

Proof. Let $\mathcal{O}$ be an open cover of $X \times Y$. For each $x \in X$, the set $\{x\} \times Y$ is compact in $X \times Y$. Therefore a finite subcollection $\mathcal{O}_x$ of $\mathcal{O}$ covers $\{x\} \times Y$. Let U_x be the union of the sets in $\mathcal{O}_x$. The set U_x is open in $X \times Y$ and contains $\{x\} \times Y$. By Lemma 7.9, for each $x \in X$ there exists an open set $W_x \subset X$ such that $x \in W_x$ and $W_x \times Y \subset U_x$. Note that $\mathcal{O}_x$ covers $W_x \times Y$.

The collection $\mathcal{W} = \{W_x \mid x \in X\}$ is an open cover of X. Since X is compact, it is covered by finitely many of the sets in $\mathcal{W}$, say $W_{x_1}, \ldots, W_{x_m}$. It follows that $\mathcal{C} = \mathcal{O}_{x_1} \cup \ldots \cup \mathcal{O}_{x_m}$ covers $X \times Y$. The collection $\mathcal{C}$ is a subcollection of $\mathcal{O}$ and is finite, being a finite union of finite sets. Therefore $\mathcal{O}$ has a finite subcollection that covers $X \times Y$, implying that $X \times Y$ is compact. ∎

Recall from Theorem 3.9 that if A and B are subsets of topological spaces X and Y, respectively, then the product of the subspace topologies that they inherit from X and Y is the same as the subspace topology that $A \times B$ inherits from $X \times Y$. In particular, Theorems 3.9 and 7.10 together imply that if A is a compact subset of X, and B is a compact subset of Y, then $A \times B$ is a compact subset of $X \times Y$. Using this result and induction, we then obtain the following corollary:

COROLLARY 7.11. *Let $X_1, \ldots, X_n$ be topological spaces, and let A_i be a compact subset of X_i for each $i = 1, \ldots, n$. Then $A_1 \times \ldots \times A_n$ is a compact subset of the product space $X_1 \times \ldots \times X_n$.*

Exercises for Section 7.1

7.1. Show that every set $A \subset \mathbb{R}$ is a compact subset of $\mathbb{R}$ in the finite complement topology on $\mathbb{R}$.

7.2. **Prove Theorem 7.6:** Let X be a topological space.
   (a) If $C_1, \ldots, C_n$ are each compact in X, then $\bigcup_{j=1}^{n} C_j$ is compact in X.
   (b) If X is Hausdorff, and $\{C_\alpha\}_{\alpha \in A}$ is a collection of sets that are compact in X, then $\bigcap_{\alpha \in A} C_\alpha$ is compact in X.

7.3. Provide an example demonstrating that an arbitrary union of compact sets in a topological space X is not necessarily compact.

7.4. Prove that $\mathbb{Z}$ with the digital line topology is not compact.

7.5. Let $A = D - \{O\}$, the disk in $\mathbb{R}^2$ with the origin removed. Show that A is not compact.

7.6. Consider the topology on $\mathbb{Z}$ generated by the basis $\mathcal{B} = \{(-n, n) \mid n \in \mathbb{Z}_+\}$.
   (a) Determine whether or not $(-5, 5)$ is a compact subset of $\mathbb{Z}$ in this topology.
   (b) Determine whether or not $\mathbb{Z}$ is compact in this topology.

7.7. Recall that the arithmetic progression topology on $\mathbb{Z}$ is generated by the basis $\mathcal{B} = \{A_{a,b} \mid a, b \in \mathbb{Z}, \ b \neq 0\}$, where each

$$A_{a,b} = \{\ldots, a - 2b, a - b, a, a + b, a + 2b, \ldots\}$$

is an arithmetic progression. Determine whether or not $\mathbb{Z}$ is compact in this topology.

7.8. Let X be a compact topological space, and let $\{C_i\}_{i \in \mathbb{Z}_+}$ be a collection of nonempty closed sets in X satisfying $C_{i+1} \subset C_i$ for each $i \in \mathbb{Z}_+$. Prove that $\bigcap_{i=1}^{\infty} C_i \neq \varnothing$.

7.9. Prove that if X is compact and Hausdorff, then X is normal. (Hint: First prove that X is regular.)

7.10. Prove the converse of Theorem 7.10: If $X \times Y$ is compact, then so are X and Y.

7.11. (a) Let $f : X \to Y$ be a continuous bijective function. Prove that if X is compact and Y is Hausdorff, then f is a homeomorphism.

   (b) Provide an example where $f : X \to Y$ is a continuous bijective function and X is compact, but f is not a homeomorphism.

   (c) Provide an example where $f : X \to Y$ is a continuous bijective function and Y is Hausdorff, but f is not a homeomorphism.

7.12. Show that the Tube Lemma does not necessarily hold if we drop the assumption that Y is compact. That is, provide an example of a noncompact space Y and an open set U in $X \times Y$ such that U contains a slice $\{x\} \times Y \subset X \times Y$ but does not contain an open tube $W \times Y$ containing the slice.

7.13. Let $f : X \to Y$ be a function. The graph of f is the subset of $X \times Y$ given by $G = \{(x, f(x)) \mid x \in X\}$.

   (a) In Exercise 4.10 we established that if $f : X \to Y$ is continuous and Y is Hausdorff, then G is a closed subset of $X \times Y$. Here we prove a converse. Assume that X and Y are topological spaces and Y is compact. Show that if G is a closed subset of $X \times Y$, then f is continuous. (Hint: Given U open in Y, and $x \in f^{-1}(U)$, show that the slice $\{x\} \times Y$ is contained in the open set $(X - G) \cup (X \times U)$, and then apply the Tube Lemma.)

   (b) Show that the result from part (a) does not hold if we drop the assumption that Y is compact. That is, find an example of a noncompact space Y and a function $f : X \to Y$ such that the graph of f is a closed subset of $X \times Y$, but f is not continuous.

7.2 *Compactness in Metric Spaces*

In real analysis, where the focus is on $\mathbb{R}^n$ with the standard metric and topology, sometimes a set is defined to be compact if it is closed and bounded. In this section, we show that in $\mathbb{R}^n$ such a definition is consistent with the topological definition already presented. We then address the extent to which this equivalence carries over to general metric spaces. Following that, we present some important convergence properties of compact sets in a metric space.

We begin with the following lemma, which we subsequently use to show that closed and bounded intervals in $\mathbb{R}$ are compact:

LEMMA 7.12. *The Nested Intervals Lemma.* *Let $\{[a_n, b_n]\}_{n \in \mathbb{Z}_+}$ be a collection of nonempty closed bounded intervals in $\mathbb{R}$ such that $[a_{n+1}, b_{n+1}] \subset [a_n, b_n]$ for each $n \in \mathbb{Z}_+$. Then $\bigcap_{n=1}^{\infty}[a_n, b_n]$ is nonempty.*

Proof. Assume that $[a_{n+1}, b_{n+1}] \subset [a_n, b_n]$ for each $n \in \mathbb{Z}_+$, as illustrated in Figure 7.8.

It follows that the interval endpoints satisfy the inequalities,

$$a_1 \leq a_2 \leq \ldots \leq a_n \leq \ldots \leq b_n \leq \ldots \leq b_2 \leq b_1.$$

The set $\{a_n\}_{n \in \mathbb{Z}_+}$ is bounded from above (for example, by each b_n) and therefore has a least upper bound A. Similarly, the set $\{b_n\}_{n \in \mathbb{Z}_+}$ has a
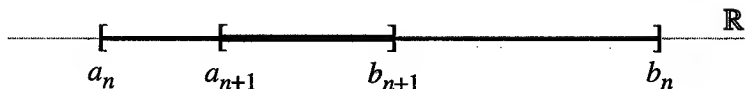
FIGURE 7.8: The interval $[a_{n+1}, b_{n+1}]$ is a subset of $[a_n, b_n]$ for each n.

greatest lower bound B. Note that $A \leq B$ and therefore the interval $[A, B]$ is nonempty.

We claim that $\bigcap_{n=1}^{\infty}[a_n, b_n] = [A, B]$, to complete the proof of the lemma. First, we show that $\bigcap_{n=1}^{\infty}[a_n, b_n] \subset [A, B]$. Thus, let $x \in \bigcap_{n=1}^{\infty}[a_n, b_n]$ be arbitrary. Then $x \in [a_n, b_n]$ for all n, implying that $x \geq a_n$ and $x \leq b_n$ for all n. Therefore $x \geq A$ and $x \leq B$; that is, $x \in [A, B]$. Hence $\bigcap_{n=1}^{\infty}[a_n, b_n] \subset [A, B]$.

To prove that $[A, B] \subset \bigcap_{n=1}^{\infty}[a_n, b_n]$, let $x \in [A, B]$ be arbitrary. Then $x \geq a_n$ and $x \leq b_n$ for all n. Therefore $x \in \bigcap_{n=1}^{\infty}[a_n, b_n]$. Thus $[A, B] \subset \bigcap_{n=1}^{\infty}[a_n, b_n]$, and it follows that $\bigcap_{n=1}^{\infty}[a_n, b_n] = [A, B]$. ∎

Lemma 7.12 does not hold if we replace the closed intervals with open intervals. For instance, the collection of nonempty bounded open intervals $\{(0, \frac{1}{n}) \mid n \in \mathbb{Z}_+\}$ satisfies the condition that $(0, \frac{1}{n+1}) \subset (0, \frac{1}{n})$ for each $n \in \mathbb{Z}_+$, but $\bigcap_{n=1}^{\infty}(0, \frac{1}{n}) = \varnothing$.

THEOREM 7.13. *Every closed and bounded interval $[a, b]$ is a compact subset of $\mathbb{R}$ with the standard topology.*

Proof. Let $\mathcal{O}$ be a cover of $[a, b]$ by open sets in $\mathbb{R}$. To derive a contradiction, assume that there is no finite subcollection of $\mathcal{O}$ covering $[a, b]$.

Consider the intervals $[a, \frac{a+b}{2}]$ and $[\frac{a+b}{2}, b]$ obtained by dividing $[a, b]$ in half. The collection $\mathcal{O}$ covers both of these intervals. For at least one of the two, there is no finite subcollection of $\mathcal{O}$ that covers it (otherwise there would be a finite subcollection of $\mathcal{O}$ covering $[a, b]$, contrary to our assumption). Choose such a half of $[a, b]$, and denote it by $[a_1, b_1]$.

In a similar manner, we can choose a half of $[a_1, b_1]$ that is not covered by a finite subcollection of $\mathcal{O}$ and denote it by $[a_2, b_2]$.

We repeat this process. In other words, given $[a_n, b_n]$, a subset of $[a, b]$ that is not covered by a finite subcollection of $\mathcal{O}$, choose a half of $[a_n, b_n]$ that is not covered by a finite subcollection of $\mathcal{O}$, and denote it by $[a_{n+1}, b_{n+1}]$. (See Figure 7.9.)
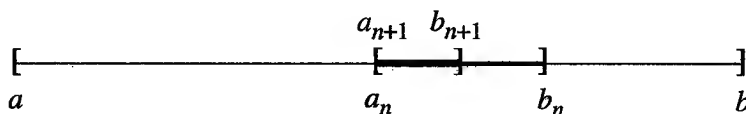
FIGURE 7.9: Each interval $[a_n, b_n]$ is not covered by a finite subcollection of $\mathcal{O}$.

Consider the collection of intervals $\{[a_n, b_n]\}_{n \in \mathbb{Z}_+}$. Based on the construction of these intervals, the following statements hold for each $n \in \mathbb{Z}_+$:

(i) $[a_{n+1}, b_{n+1}] \subset [a_n, b_n]$,

(ii) $b_n - a_n = \frac{b-a}{2^N}$,

(iii) $[a_n, b_n]$ is not covered by a finite subcollection of $\mathcal{O}$.

By Lemma 7.12, it follows that $\bigcap_{n=1}^{\infty}[a_n, b_n]$ is nonempty. Let x be in this intersection. Then $x \in [a, b]$, and therefore there exists $U \in \mathcal{O}$ such that $x \in U$. Since U is open in $\mathbb{R}$, there exists $\varepsilon > 0$ such that $(x - \varepsilon, x + \varepsilon) \subset U$. Let N be a positive integer large enough so that $\frac{b-a}{2^N} < \varepsilon$. Since $x \in \bigcap_{n=1}^{\infty}[a_n, b_n]$, it follows that $x \in [a_N, b_N]$. Furthermore, since $b_N - a_N = \frac{b-a}{2^N} < \varepsilon$, it follows that $[a_N, b_N] \subset (x - \varepsilon, x + \varepsilon) \subset U$. But then $[a_N, b_N]$ is covered by a single set in $\mathcal{O}$, contradicting the fact that $[a_N, b_N]$ is not covered by a finite subcollection of $\mathcal{O}$.

Thus there must be a finite subcollection of $\mathcal{O}$ that covers $[a, b]$, and it follows that $[a, b]$ is compact. ∎

EXAMPLE 7.6. Recall from Example 3.17 that a topological graph is a space obtained by taking finitely many points (vertices) and finitely many closed bounded intervals and gluing the endpoints of the intervals to the vertices. A collection of finitely many points and finitely many closed bounded intervals is a compact space since it is a finite union of compact spaces. Thus every topological graph is compact since it is the image of a compact space under a quotient mapping.

In the next theorem, we extend Theorem 7.13 to a corresponding result for products of closed bounded intervals in $\mathbb{R}^n$. The proof follows directly from Theorem 7.13 and Corollary 7.11.

THEOREM 7.14. *Let* $[a_1, b_1], \ldots, [a_n, b_n]$ *be closed bounded intervals in* $\mathbb{R}$. *Then* $[a_1, b_1] \times \ldots \times [a_n, b_n]$ *is a compact subset of* $\mathbb{R}^n$. *(See Figure 7.10.)*
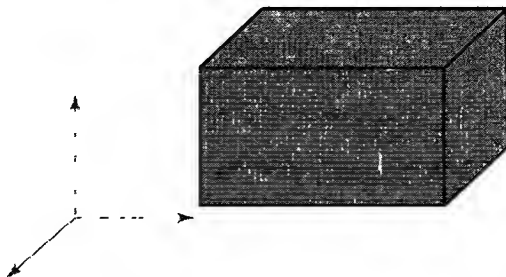

FIGURE 7.10: A product of closed bounded intervals is a compact subset of $\mathbb{R}^n$.

Now, using Theorem 7.14 we can completely determine the compact sub-sets of $\mathbb{R}^n$ in the standard topology. Recall that in a metric space (X, d) a set A is said to be bounded if there exists $m > 0$ such that $d(x, y) \leq m$ for all $x, y \in A$.

THEOREM 7.15. *Let $\mathbb{R}^n$ have the standard topology and the standard metric d. A set $A \subset \mathbb{R}^n$ is compact in $\mathbb{R}^n$ if and only if it is closed and bounded.*

Proof. Let A be compact in $\mathbb{R}^n$. Then since $\mathbb{R}^n$ is Hausdorff, it follows by Theorem 7.8 that A is closed. To see that A is bounded, consider the collection $\mathcal{O} = \{B(O, n) \mid n \in \mathbb{Z}_+\}$, made up of open balls centered at the origin in $\mathbb{R}^n$. The collection $\mathcal{O}$ is an open cover of A, and since A is compact, it follows that finitely many of the sets in $\mathcal{O}$ cover A. Thus, there exists $N \in \mathbb{Z}_+$ such that $A \subset B(O, N)$. Therefore, for $x, y \in A$, we have $d(x. y) < 2N$, implying that A is bounded.

Now assume that A is closed and bounded. Let $a = (a_1, \ldots, a_n)$ be a point in A, and assume that $d(x, y) < M$ for all x and y in A. Then A is contained in the product of intervals

$$P = [a_1 - M, a_1 + M] \times \ldots \times [a_n - M, a_n + M],$$

as illustrated for the two-dimensional case in Figure 7.11. The set P is a compact subset of $\mathbb{R}^n$ by Theorem 7.14. Since A is closed and a subset of P, it follows by Theorem 7.7 that A is compact in $\mathbb{R}^n$. ∎
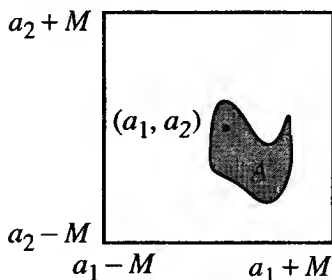
FIGURE 7.11: The set A is a subset of the compact set $[a_1 - M, a_1 + M] \times [a_2 - M, a_2 + M]$.

For subsets of $\mathbb{R}^n$ in the standard metric and topology, Theorem 7.15 implies that defining compact as closed and bounded is equivalent to defining it in terms of open coverings.

Now, again considering Figure 7.1, we see that the sets in the plane that are labeled "compact" are exactly those sets that are closed and bounded, and the sets that are labeled "noncompact" are the sets that are either not closed or not bounded.

EXAMPLE 7.7. The circle S^1 is compact since it is a closed and bounded subset of $\mathbb{R}^2$. Recall that a knot is an embedding of S^1 in $\mathbb{R}^3$. Since S^1 is compact, the image of such an embedding is compact in $\mathbb{R}^3$. We usually think of a knot as the image of an embedding in $\mathbb{R}^3$, rather than the embedding itself. Thus, we say that a knot is a compact subset of $\mathbb{R}^3$.

EXAMPLE 7.8. The torus T is compact. We prove this in three ways.

Method 1:   The torus T is the subspace of $\mathbb{R}^3$ obtained by rotating a circle about the z-axis (as described in Example 3.5). The torus is closed since given any point in its complement, there is an open ball of sufficiently small radius centered at that point and disjoint from the torus. Also, the torus is bounded since it is contained in the ball of radius 4 centered at the origin. Since the torus is closed and bounded in $\mathbb{R}^3$, it is compact.

Method 2:   The torus is homeomorphic to the product space $S^1 \times S^1$. The circle S^1 is compact. Therefore the torus is the product of two compact spaces, and thus it is compact.

Method 3:   The torus is homeomorphic to a space obtained as a quotient space of the square by identifying opposite edges. The square is a closed and bounded subset of $\mathbb{R}^2$ and therefore is compact. Furthermore, a quotient map on the square is a continuous function, and the image of a compact space under a continuous function is compact. Therefore the torus is compact.

The second method in Example 7.8 also allows us to conclude that for each $n > 2$, the n-torus is a compact topological space. Furthermore, the third method allows us to conclude that the annulus, Möbius band, Klein bottle, sphere, and projective plane are all compact spaces as well.

Now, if in Theorem 7.15 we replace $\mathbb{R}^n$ with an arbitrary metric space, is being compact still equivalent to being closed and bounded in this, more general, setting? The answer is no. One implication in the equivalence continues to work; that is, in a metric space, a compact subset is closed and bounded. However, a closed and bounded subset of a metric space need not be compact. We ask you to prove these results in Exercise 7.19.

One of the properties that mathematicians were trying to capture in the concept of compactness is that sequences have a convergent subsequence. In the next theorem we show that this holds for compact sets in a metric space.

THEOREM 7.16. *Let (X, d) be a metric space, and assume that A is compact in X. If (x_n) is a sequence in A, then there exists a subsequence (x_{n_m}) of (x_n) that converges to a limit in A.*

Proof. We claim that there exists $a \in A$ such that every nonempty open ball centered at a contains x_n for infinitely many n. Suppose this does not hold. Then for every $a \in A$ there exists ε_a such that $B_d(a, \varepsilon_a)$ contains

members of the sequence (x_n) for at most finitely many n. Consider the collection $\mathcal{O} = \{B_d(a, \varepsilon_a)\}_{a \in A}$. No finite subcollection of $\mathcal{O}$ can cover the sequence (x_n) since each set $B_d(a, \varepsilon_a)$ contains members of the sequence (x_n) for only at most finitely many n. On the other hand, $\mathcal{O}$ is an open cover of the compact set A and therefore there is a finite subcollection of $\mathcal{O}$ that covers A. Since the sequence (x_n) is in A, this is a contradiction. Thus, there exists $a \in A$ such that every nonempty open ball centered at a contains x_n for infinitely many n.

Now, for each $m \in \mathbb{Z}_+$ let x_{n_m} be a point in the sequence (x_n) such that $n_m > n_{m-1}$ and $x_{n_m} \in B_d(a, \frac{1}{m})$. Then (x_{n_m}) is a subsequence of (x_n) converging to $a \in A$. ∎

EXAMPLE 7.9. Consider the sequence in $[0, 1]$ given by

$$(\tfrac{1}{2}, \tfrac{1}{4}, \tfrac{3}{4}, \tfrac{1}{8}, \tfrac{3}{8}, \tfrac{5}{8}, \tfrac{7}{8}, \tfrac{1}{16}, \tfrac{3}{16}, \ldots).$$

The sequence bounces around in $[0, 1]$, and no term in the sequence repeats. However, since $[0, 1]$ is compact, Theorem 7.16 guarantees that there is a subsequence that converges to some point $a \in [0, 1]$. Interestingly, in this case we can do better. In fact, for every $a \in [0, 1]$ there is a subsequence converging to a. (See Exercise 7.20.)

An important convergence question for metric spaces is the following: If the points in a sequence (x_n) get closer and closer to each other, then does the sequence converge? We investigate this question in the remainder of the section.

DEFINITION 7.17. *Let (X, d) be a metric space. A sequence (x_n) in X is called a **Cauchy sequence** if for every $\varepsilon > 0$ there exists $N \in \mathbb{Z}_+$ such that $d(x_n, x_m) < \varepsilon$ for every $n, m \geq N$.*

The defining condition for a Cauchy sequence indicates that the points in the sequence (x_n) are getting close to each other as n gets large, but it does not say anything about whether or not the sequence converges. For $\mathbb{R}^n$, the following theorem settles the question about the convergence of Cauchy sequences:

THEOREM 7.18. *Let (x_n) be a Cauchy sequence in $\mathbb{R}^n$ with the standard metric d. Then (x_n) converges to a limit in $\mathbb{R}^n$.*

Proof. Suppose that (x_n) is a Cauchy sequence in $\mathbb{R}^n$. Pick a $\mu >$ 0. Since (x_n) is a Cauchy sequence, there exists $N \in \mathbb{Z}_+$ such that $d(x_n, x_m) < \mu$ for all $n, m \geq N$. Let C be the closed ball of radius μ centered at x_N. Then C is a compact set since it is closed and bounded. Also, the sequence $(x_N, x_{N+1}, \ldots)$ is in C. By Theorem 7.16, the sequence $(x_N, x_{N+1}, \ldots)$ has a subsequence that converges to a limit x^* in C. We claim that the original sequence (x_n) converges to x^*.

Thus, suppose that $\varepsilon > 0$ is arbitrary. Because (x_n) is a Cauchy sequence, we can choose $M \in \mathbb{Z}_+$ such that $d(x_n, x_m) < \varepsilon/2$ for every $n, m \geq M$. Now, let $k \geq M$ be arbitrary. We show that $d(x_k, x^*) < \varepsilon$, proving that (x_n) converges to x^*. Since $(x_N, x_{N+1}, \ldots)$ has a subsequence converging to x^*, there exists $j \geq M$ such that $d(x_j, x^*) < \varepsilon/2$. Also, having $j, k \geq M$ implies that $d(x_k, x_j) < \varepsilon/2$. Therefore, by the triangle inequality, it follows that $d(x_k, x^*) < \varepsilon$, as desired. ∎

DEFINITION 7.19. *A metric space X is called **complete** if every Cauchy sequence in X converges to a limit in X.*

By Theorem 7.18, it follows that $\mathbb{R}^n$ with the standard metric is a complete metric space. However, if we let $X = \mathbb{R} - \{0\}$ with the metric $d(x, y) = |x - y|$, then X is not a complete metric space. The sequence $(1, \frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \ldots)$ is a Cauchy sequence in X that does not converge in X.

The following theorem is another useful convergence result for compact metric spaces.

THEOREM 7.20. *If X is a compact metric space, then X is complete.*

Proof. See Exercise 7.21. ∎

Exercises for Section 7.2

7.14. Does Lemma 7.12 hold if we replace nonempty closed bounded intervals in $\mathbb{R}$ with nonempty closed sets in $\mathbb{R}$? Specifically, prove the following implication or find a counterexample: If $\{A_n\}_{n \in \mathbb{Z}_+}$ is a collection of nonempty closed sets in $\mathbb{R}$ such that $A_{n+1} \subset A_n$ for each $n \in \mathbb{Z}_+$, then $\bigcap_{n \in \mathbb{Z}_+} A_n$ is nonempty.

7.15. Show that $[0, 1]$ is not a compact subset of $\mathbb{R}$ with the lower limit topology.

7.16. If $\mathbb{R}_d$ denotes $\mathbb{R}$ with the discrete topology, determine which of the subsets, $A = \{(x, 2) \mid 1 \leq x \leq 2\}$ and $B = \{(3, y) \mid 1 \leq y \leq 2\}$, as shown in Figure 7.12, is a compact subset of $\mathbb{R}_d \times \mathbb{R}$.

7.17. Use compactness to prove that the plane is not homeomorphic to the sphere. (Recall, in Section 6.2 we distinguished between a number of pairs of spaces, including the line and the plane and the line and the sphere, but we indicated that we were not yet in a position to distinguish between the plane and the sphere. With compactness, we can now make that distinction.)
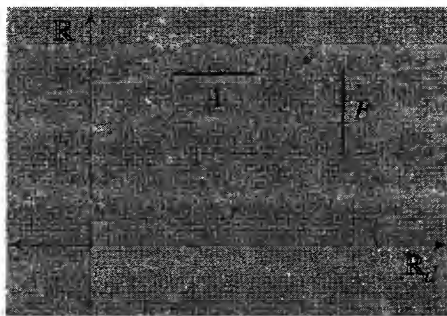
FIGURE 7.12: Is either A or B a compact subset of $\mathbb{R}_d \times \mathbb{R}$?

7.18. In this exercise we demonstrate that if we drop the condition that X is Hausdorff in Theorem 7.6, then the intersection of compact sets in X is not necessarily a compact set. Define the **extra-point line** as follows. Let $X = \mathbb{R} \cup \{p_e\}$, where p_e is an extra point, not contained in $\mathbb{R}$. Let $\mathcal{B}$ be the collection of subsets of X consisting of all intervals $(a, b) \subset \mathbb{R}$ and all sets of the form $(c, 0) \cup \{p_e\} \cup (0, d)$ for $c < 0$ and $d > 0$.
   (a) Prove that $\mathcal{B}$ is a basis for a topology on X.
   (b) Show that the resulting topology on X is not Hausdorff.
   (c) Find two compact subsets of X whose intersection is not compact. Prove that the sets are compact and that the intersection is not.

7.19. (a) Let (X, d) be a metric space. Prove that if A is compact in X, then A is closed in X and bounded under the metric d.
   (b) Provide an example demonstrating that a subset of a metric space can be closed and bounded but not compact.

7.20. Consider the sequence defined in Example 7.9. Prove that for every $a \in [0, 1]$ there is a subsequence converging to a.

7.21. Prove Theorem 7.20: If X is a compact metric space, then X is complete.

7.3   The Extreme Value Theorem

The Extreme Value Theorem, like the Intermediate Value Theorem, is a topologically based theorem that is often introduced in a calculus course. As we saw in Chapter 6, the Intermediate Value Theorem concerns continuous real-valued functions on a connected domain. The Extreme Value Theorem concerns continuous real-valued functions on a compact domain. In this section we prove a general version of the Extreme Value Theorem and consider a number of its useful consequences.

To begin, we need the following lemma, which indicate that every compact subset of the real line contains a maximum value and a minimum value. (See Figure 7.13.)

FIGURE 7.13: Every compact subset of $\mathbb{R}$ has maximum and minimum values.

LEMMA 7.21. *Let A be a compact subset of $\mathbb{R}$. Then there exist $m, M \in A$ such that $m \leq a \leq M$ for all $a \in A$.*

> **Proof.** Here we prove the existence of the maximum value M. The proof of the existence of the minimum value m is similar. Since A is compact, it is closed and bounded in $\mathbb{R}$. Therefore A is bounded from above. It follows that the set A has a least upper bound; denote it by M. Of course, $a \leq M$ for all $a \in A$. We claim that $M \in A$. We prove the claim by contradiction; thus suppose that $M \notin A$. Since A is closed, it follows that there exists an $\varepsilon > 0$ such that $(M - \varepsilon, M + \varepsilon) \cap A = \varnothing$. Then $M - \frac{\varepsilon}{2}$ is an upper bound for A that is smaller than M, a contradiction. Therefore $M \in A$, and A has a maximum value. ∎

Now, using Lemma 7.21, we establish the Extreme Value Theorem.

THEOREM 7.22. *The Extreme Value Theorem (General Version).* *Let X be compact and $f : X \to \mathbb{R}$ be continuous. Then f takes on a maximum value and a minimum value on X; that is, there exist $a, b \in X$ such that $f(a) \leq f(x) \leq f(b)$ for all $x \in X$.*

> **Proof.** Since X is compact and f is continuous, $f(X)$ is a compact subset of $\mathbb{R}$ by Theorem 7.5. Therefore $f(X)$ contains a maximum value M and a minimum value m by Lemma 7.21. The points m and M are in $f(X)$; therefore there exist $a, b \in X$ such that $f(a) = m$ and $f(b) = M$. Now, for all $x \in X$ we have that $f(x) \in f(X)$, and thus $f(a) = m \leq f(x) \leq M = f(b)$, as we wished to show. ∎

EXAMPLE 7.10. In this simple application of the Extreme Value Theorem, we view the surface of the Earth as a sphere and the surface temperature as a continuous function on the sphere. Since the sphere is compact, the Extreme Value Theorem ensures that somewhere on the Earth there is a point or there are points where the temperature is hotter than those at every other point on Earth, and there are also points where the temperature is colder than those at every other point on Earth.

The following corollary of Theorem 7.22 is the version of the Extreme Value Theorem that is usually encountered in a calculus course. It follows from Theorem 7.22 if we let $X = [a, b]$.

COROLLARY 7.23. *The Extreme Value Theorem on [a, b].* *Assume that $f : [a, b] \to \mathbb{R}$ is continuous. Then f takes on a maximum value and a minimum value on $[a, b]$.*

Combining the Extreme Value Theorem and the Intermediate Value Theorem, we obtain the following corollary:

COROLLARY 7.24. *Let $[a, b]$ be a closed and bounded interval in $\mathbb{R}$, and assume that $f : [a, b] \to \mathbb{R}$ is continuous. Then the image of f is a closed and bounded interval in $\mathbb{R}$. (See Figure 7.14.)*
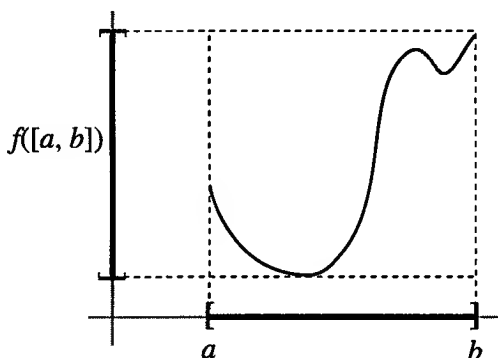
Proof. See Exercise 7.22. ∎

FIGURE 7.14: For a continuous $f : [a, b] \to \mathbb{R}$, the image $f([a, b])$ is a closed and bounded interval in $\mathbb{R}$.

The Extreme Value Theorem is the basis for a variety of optimization theorems and applications. Consider the following example:

EXAMPLE 7.11. The publishers at Klein's Publishing House would like to catch the new wave of interest in applied topology. They are planning to publish an applied topology text to compete with those currently on the market. The business department has committed a $500,000 budget to the first run of the book. The budget is to be allocated among several aspects of the book's production, including an authorship contract, editorial costs, printing, advertising, and distribution. Suppose that there are n such variables, $v_1, \ldots, v_n$. The profit that the publishers can expect to make on this venture depends on how the resources are allocated. For example, if they opt for printing the book in a high-cost Möbius-band format and a minimal advertising budget, they will realize a smaller profit than if they had used a standard book format and a larger advertising budget. Thus, we regard the profit, P, as a function of the variables $v_1, \ldots, v_n$. It is natural to assume that P is continuous. The domain of P is the subset of $\mathbb{R}^n$ given by

$$D = \{(v_1, \ldots, v_n) \mid v_1 \geq 0, \ldots, v_n \geq 0, \ v_1 + \ldots + v_n \leq 500{,}000\}.$$

Since D is closed and bounded, it is compact. Therefore the Extreme Value Theorem indicates that there is a choice for the allocation of resources resulting in a maximum profit for the planned textbook project.

While the Extreme Value Theorem does not tell us how to find the maximum and minimum values of a function, it guarantees that they exist. This is typical of this kind of topology theorem. Like the Intermediate Value Theorem, the Extreme Value Theorem asserts the existence of a point with specified properties, but does not provide an exact location for it.

The next two theorems will be helpful to us in establishing subsequent results in the text. The Extreme Value Theorem is used in the proof of each.

Recall that if A and B are subsets of a metric space (X, d), then we define the distance between A and B by

$$d(A, B) = \text{glb}\{d(a, b) \mid a \in A, \ b \in B\}.$$

The following theorem indicates that there is a positive distance between two disjoint compact sets in a metric space.

THEOREM 7.25. *Let (X, d) be a metric space. If A and B are disjoint compact subsets of X, then $d(A, B) > 0$.*

> **Proof.** The distance function $d : X \times X \to \mathbb{R}$ is continuous (see Exercise 5.13), and the set $A \times B$ is compact in $X \times X$ by Corollary 7.11. Therefore the Extreme Value Theorem applies and indicates that d takes on a minimum value on $A \times B$. That is, there exist $a^* \in A$ and $b^* \in B$ such that $d(a^*, b^*) \leq d(a, b)$ for every $a \in A$ and $b \in B$. It follows that $d(A, B) = d(a^*, b^*)$. Since A and B are disjoint, $a^* \neq b^*$, and therefore $d(a^*, b^*) > 0$. Hence, $d(A, B) > 0$, as desired.   ∎

IMPORTANT NOTE: *The result of Theorem 7.25 does not hold if we replace "compact" with "closed." That is, in a metric space it is possible to have disjoint closed sets A and B with $d(A, B) = 0$. (See Exercise 7.23.)*

In the next results, we work with functions defined by taking the distance between a point and a set in a metric space. Specifically, let X be a metric space and let A be a subset of X. Define $f_A : X \to \mathbb{R}$ by $f_A(x) = d(\{x\}, A)$. Thus, $f_A(x)$ is the distance from the single-point set $\{x\}$ to the set A. We can regard it as the distance from the point x to the set A.

LEMMA 7.26. *Let (X, d) be a metric space, and let A be a subset of X. The function $f_A : X \to \mathbb{R}$, defined by $f_A(x) = d(\{x\}, A)$, is continuous.*

> **Proof.** See Exercise 7.24.   ∎

The following lemma tells us that, given an open cover of a compact metric space, there is a threshold value such that every open ball with a radius below the threshold is guaranteed to lie in some set in the open cover.

LEMMA 7.27. *The Lebesgue Number Lemma.* *Let (X, d) be a compact metric space, and let $\mathcal{O}$ be an open cover of X. Then there is a $\lambda > 0$ such that for every $x \in X$ there exists a $U \in \mathcal{O}$ satisfying $B_d(x, \lambda) \subset U$.*

The number λ is called a **Lebesgue number** for the cover $\mathcal{O}$.

Proof. Let $\mathcal{O}$ be an open cover of X. If the whole space X is in the cover, then any real number λ serves as a Lebesgue number for the cover. Thus assume X is not one of the sets in the cover.

Since X is compact, finitely many of the sets from $\mathcal{O}$ cover X, say $U_1, \ldots, U_n$. For each $i = 1, \ldots, n$, set $C_i = X - U_i$. Each set C_i is nonempty since we are assuming that none of the sets in $\mathcal{O}$ is equal to X. Define $f : X \to \mathbb{R}$ by

$$f(x) = \frac{1}{n} \sum_{i=1}^{n} f_{C_i}(x).$$

Thus, $f(x)$ is the average of the distances from x to each set C_i.

We claim that $f(x) > 0$ for all $x \in X$. To prove the claim, let $x \in X$ be arbitrary. Since the sets U_i cover X, there exists k such that $x \in U_k$. The set U_k is open; consequently there exists $\varepsilon > 0$ such that $B_d(x, \varepsilon) \in U_k$. It follows that $f_{C_k}(x) > \varepsilon$, and therefore $f(x) > \varepsilon/n$.

The function f is continuous since each f_{C_i} is continuous by Lemma 7.26 and since addition and multiplication of continuous functions is continuous. (See Exercise 4.16.) Furthermore, the domain of f is compact. Therefore the Extreme Value Theorem implies that f takes on a minimum value λ. Since $f(x) > 0$ for all x, it must be that $\lambda > 0$.

We show that λ is the desired Lebesgue number. Thus, let $x \in X$ be arbitrary, and consider $B_d(x, \lambda)$. We claim that this open ball is a subset of at least one of the sets U_i and therefore is a subset of some set from the cover $\mathcal{O}$. Suppose none of the sets U_i contains $B_d(x, \lambda)$. Then $f_{C_i}(x) < \lambda$ for all i, implying that $f(x) < \lambda$. But this contradicts the fact that λ is the absolute minimum of f over X. It follows that λ satisfies the requirements of the lemma. ∎

We will find the following corollary of Lemma 7.27 helpful in some of the work we do in the remainder of the text.

COROLLARY 7.28. *Let $\mathcal{O}$ be a cover of the closed bounded interval $[a, b]$ by sets that are open in $\mathbb{R}$. Then there is a subdivision*

$$a = a_0 < a_1 < \ldots < a_n = b$$

of $[a, b]$ such that for all $j = 1, \ldots, n$, there exists a $U_j \in \mathcal{O}$ containing $[a_{j-1}, a_j]$.

Proof. See Exercise 7.26. ∎

The Fundamental Theorem of Calculus is another theorem whose proof uses the Extreme Value Theorem. Specifically, the Extreme Value Theorem is used to prove Rolle's Theorem, which is equivalent to the Mean Value Theorem, and the Mean Value Theorem is an important tool employed in proving the Fundamental Theorem of Calculus.

In Chapter 9 we use topological tools to prove the Fundamental Theorem of Algebra. We do not go so far as to say that the Fundamental Theorem of

Calculus and the Fundamental Theorem of Algebra are theorems of topology. What is important to realize, however, is how intertwined the disciplines of algebra, analysis (of which calculus is a part), and topology are. Although typically they are distinct courses at the introductory level in an abstract mathematics program, they are far from distinct in the topics, tools, and properties they cover. This should not be a surprise, because these disciplines have all arisen from a desire to understand the structure of the real-number system, Euclidean space, and functions between them.

Exercises for Section 7.3

7.22. **Prove Corollary 7.24:** Let $[a, b]$ be a closed and bounded interval in $\mathbb{R}$, and assume that $f : [a, b] \to \mathbb{R}$ is continuous. Then the image of f is a closed and bounded interval in $\mathbb{R}$.

7.23. Provide an example of closed sets, A and B, in a metric space (X, d) such that A and B are disjoint and $d(A, B) = 0$.

7.24. **Prove Lemma 7.26:** Let (X, d) be a metric space, and let A be a subset of X. The function $f_A : X \to \mathbb{R}$, defined by $f_A(x) = d(\{x\}, A)$, is continuous.

7.25. Provide an example showing that the Lebesgue Number Lemma does not hold if we drop the assumption that the space is compact.

7.26. **Prove Corollary 7.28:** Let $\mathcal{O}$ be a cover of the closed bounded interval $[a, b]$ by sets that are open in $\mathbb{R}$. Then there is a subdivision $a = a_0 < a_1 < \ldots < a_n = b$ of $[a, b]$ such that for all $j = 1, \ldots, n$, there exists a $U_j \in \mathcal{O}$ containing $[a_{j-1}, a_j]$.

Supplementary Exercises: The Tietze Extension Theorem

Let A be a subset of a topological space X, and let $f : A \to Y$ be continuous. It is often of interest to know if f extends to a continuous function $F : X \to Y$; that is, if there exists a continuous $F : X \to Y$ such that $F(a) = f(a)$ for all $a \in A$.

The Tietze Extension Theorem provides general conditions under which it can be concluded that extensions exist. Specifically, if A is a closed subset of a normal space X, and $J \subset \mathbb{R}$ is either a closed bounded interval, an open interval, or all of $\mathbb{R}$, then the Tietze Extension Theorem asserts that every continuous $f : A \to J$ extends to a continuous function $F : X \to J$.

In these exercises, we work through a proof of the following special case of the Tietze Extension Theorem. A proof of the general Tietze Extension Theorem can be found in [Mun].

THEOREM 7.29. *The Tietze Extension Theorem for Metric Spaces.* *Let A be a closed subset of a metric space X; then every continuous $f : A \to [-1, 1]$ extends to a continuous function $F : X \to [-1, 1]$.*

Our proof of the Tietze Extension Theorem uses the Uniform Convergence Theorem (Theorem 4.13), which was established in a set of supplementary exercises in Section 4.1.

We use the Tietze Extension Theorem to prove a retraction existence theorem (Theorem 9.14) in Chapter 9. That result is then used in the proof of the Jordan Curve Theorem in Chapter 11.

Before proving Theorem 7.29, we need a few supporting results. The following lemma indicates that if B and C are disjoint nonempty closed subsets of a metric space X, then there exists a continuous real-valued function that takes on constant values b and c on B and C, respectively, and takes on values between b and c on the complement of $B \cup C$:

LEMMA 7.30. *Let B and C be disjoint nonempty closed subsets of a metric space X, and assume $b, c \in \mathbb{R}$ with $b < c$. Then there exists a continuous function $g : X \to [b, c]$ such that $B = g^{-1}(b)$ and $C = g^{-1}(c)$.*

SE 7.27. As in Lemma 7.26, let $f_B(x)$ and $f_C(x)$ be the distances from $x \in X$ to the sets B and C, respectively. Show that the requirements of Lemma 7.30 are satisfied by the function $g : X \to [b, c]$, defined by

$$g(x) = \frac{cf_B(x) + bf_C(x)}{f_B(x) + f_C(x)}.$$

With Lemma 7.30 we can prove the following lemma, which we then use in the proof of the Tietze Extension Theorem:

LEMMA 7.31. *Let X be a metric space, and let $A \subset X$ be closed. Assume that $f : A \to [-k, k]$ is continuous. Then there exists a continuous $g : X \to [-\frac{k}{3}, \frac{k}{3}]$ such that $|f(a) - g(a)| \leq \frac{2k}{3}$ for all $a \in A$.*

The idea in proving Lemma 7.31 is to have $g(x)$ equal $-\frac{k}{3}$ for those x for which $f(x) \leq -\frac{k}{3}$, to have $g(x)$ equal $\frac{k}{3}$ for those x for which $f(x) \geq \frac{k}{3}$, and to otherwise have g continuously transition between $-\frac{k}{3}$ and $\frac{k}{3}$. (See Figure 7.15.) Specifically, let $B = f^{-1}([-k, -\frac{k}{3}])$ and $C = f^{-1}([\frac{k}{3}, k])$. The sets B and C are disjoint closed subsets of A, and since A is closed in X, they are also closed in X. By Lemma 7.30, there exists a continuous function $g : X \to [-\frac{k}{3}, \frac{k}{3}]$ such that $g^{-1}(-\frac{k}{3}) = B$ and $g^{-1}(\frac{k}{3}) = C$.
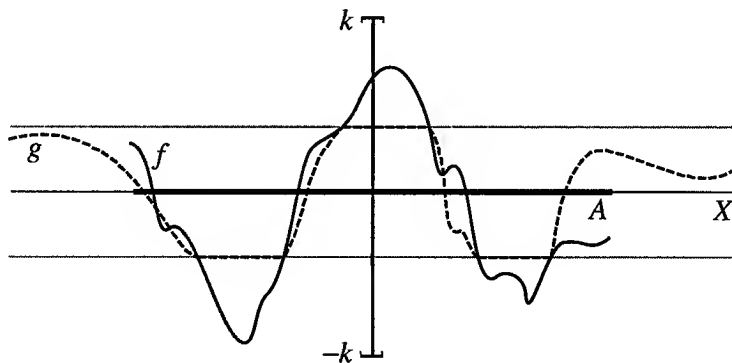

FIGURE 7.15: The function g approximates f and maps into $[-\frac{k}{3}, \frac{k}{3}]$.

SE 7.28. Prove that $|f(a) - g(a)| \leq \frac{2k}{3}$ for all $a \in A$.

Proof of the Tietze Extension Theorem for Metric Spaces. Let A be a closed subset of a metric space X, and let $f : A \to [-1, 1]$ be continuous. To prove the theorem, we build a sequence of continuous functions F_n on X that, on A, are progressively better approximations to f. In the limit, we obtain a continuous function F on X that is equal to f on A.

By Lemma 7.31, there exists a $g_1 : X \to [-\frac{1}{3}, \frac{1}{3}]$ such that $|f(x) - g_1(x)| \le \frac{2}{3}$ for all $x \in A$. Now, the continuous function $f - g_1$ maps A into the interval $[-\frac{2}{3}, \frac{2}{3}]$. Using Lemma 7.31 again, there exists a $g_2 : X \to [-\frac{2}{9}, \frac{2}{9}]$ such that $|f(x) - g_1(x) - g_2(x)| \le \frac{4}{9}$. Continuing this process, we obtain a sequence of continuous functions

$$g_n : X \to \left[\frac{-2^{n-1}}{3^n}, \frac{2^{n-1}}{3^n} \right]$$

such that $|f(a) - g_1(a) - \ldots - g_n(a)| \le (\frac{2}{3})^n$ for all $a \in A$.

SE 7.29. Prove that $|g_1(x) + \ldots + g_n(x)| \le 1$ for all $x \in X$.

For each $n \in \mathbb{Z}_+$, define $F_n : X \to [-1, 1]$ by $F_n(x) = g_1(x) + \ldots + g_n(x)$. Each function F_n is continuous, being a sum of continuous functions.

SE 7.30. Prove that for each $x \in X$ the sequence of function values $(F_n(x))$ is a Cauchy sequence in $[-1, 1]$.

Since $(F_n(x))$ is a Cauchy sequence in $[-1, 1]$ for each $x \in X$, Theorem 7.20 implies that the sequence $(F_n(x))$ converges to a limit $F(x) \in [-1, 1]$. In this way, a function $F : X \to [-1, 1]$ is defined. We claim that F is the desired extension of f. First we show that F is continuous.

SE 7.31. Prove that the sequence of functions F_n converges uniformly to the function F.

By the Uniform Convergence Theorem (Theorem 4.13), it follows that F is continuous.

SE 7.32. Prove that F is an extension of f; that is, show that $F(a) = f(a)$ for every $a \in A$.

Consequently, F is the desired continuous extension of f, completing the proof of the Tietze Extension Theorem. ∎

7.4   Limit Point Compactness

Before mathematicians settled on the definition of compactness we presented in Section 7.1, other definitions were considered. For example, there was the following candidate:

DEFINITION 7.32. *A topological space X is **limit point compact** if every infinite subset of X has a limit point.*

EXAMPLE 7.12. In $\mathbb{R}$, the subspace $A = \{0\} \cup \{\frac{1}{n} \mid n \in \mathbb{Z}_+\}$ is limit point compact. Let B be an infinite subset of A. Then B must contain values of the form $1/n$ where n is arbitrarily large. Therefore 0 is a limit point of B. Thus, every infinite subset of A has a limit point, implying that A is limit point compact.

EXAMPLE 7.13. Let X be an infinite set with the finite complement topology. We show that X is limit point compact. Thus, let B be an infinite subset of X. We claim that every point of X is a limit point of B. If $x \in X$, and U is a neighborhood of x, then since U contains all but finitely many points of X, it intersects B in infinitely many points. In particular, U intersects B in points other than x, implying that x is a limit point of B. Therefore every infinite subset B of X has a limit point, implying that X is limit point compact.

Examples 7.3 and 7.12 together establish that in $\mathbb{R}$, the subspace $A = \{0\} \cup \{\frac{1}{n} \mid n \in \mathbb{Z}_+\}$ is both compact and limit point compact. The following theorem indicates that this is no coincidence:

THEOREM 7.33. *If a topological space X is compact, then it is limit point compact.*

> *Proof.* We prove that if X is not limit point compact, then X is not compact. Suppose that X is not limit point compact. Thus, X has an infinite subset B that does not have a limit point. Since each $x \in B$ is not a limit point of B, it follows that for every $x \in B$ there exists a neighborhood U_x of x such that $U_x \cap B = \{x\}$. Also, since B has no limit points, Corollary 2.9 implies that B is closed. Hence, $X - B$ is open. Let $\mathcal{O}$ be the collection of open sets $\{U_x \mid x \in B\} \cup \{X - B\}$. Then $\mathcal{O}$ is an open cover of X. Furthermore $\mathcal{O}$ has no finite subcover, since each of the infinitely many points x in B is contained in only the open set U_x in $\mathcal{O}$. Therefore X has an open cover with no finite subcover. It follows that X is not compact. ∎

EXAMPLE 7.14. The disk is compact and therefore limit point compact. This implies that every infinite subset has a limit point. However, there are infinite subsets that do not have obvious limit points. For example, take the set $A \subset \mathbb{R}^2$, expressed in (r, θ) polar coordinates as

$$A = \left\{ \left(\frac{\theta - 1}{\theta}, \theta \right) \, \middle| \, \theta = 1, 2, 3, \dots \right\}$$

and illustrated in Figure 7.16. This set of points spirals out toward the circle S^1, but there is no obvious point where this set accumulates. Even so, Theorem 7.33 tells us that this set must have at least one limit point.
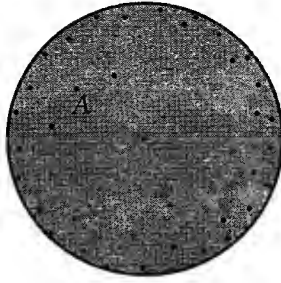
FIGURE 7.16: There must be a limit point of A in the disk.

The converse of Theorem 7.33 does not hold. As the following example demonstrates, limit point compactness does not necessarily imply compactness.

EXAMPLE 7.15. Let $\mathbb{Z}$ have the topology generated by the basis

$$\mathcal{B} = \{(-n, n) \mid n \in \mathbb{Z}_+\}.$$

In this topology, $\mathbb{Z}$ is not compact since the basis $\mathcal{B}$ is an open cover that has no finite subcover. We claim, however, that $\mathbb{Z}$ is limit point compact in this topology. In fact, we show that every nonempty subset of $\mathbb{Z}$ has a limit point. Thus, let A be a nonempty subset of $\mathbb{Z}$. Every subset of $\mathbb{Z}$ has an element with minimum absolute value; let s be such an element of A. For every $t \in \mathbb{Z}$ such that $|t| > |s|$, it follows that every open set containing t must contain s. Therefore every t satisfying $|t| > |s|$ is a limit point of A. Hence, A has a limit point, implying that $\mathbb{Z}$ is limit point compact in the topology generated by $\mathcal{B}$.

If we restrict ourselves to metric spaces, then limit point compactness and compactness are equivalent. This is yet one more example of the convenient properties that hold in a metric space. We present this result in Theorem 7.36, but before doing so we prove a couple of helpful lemmas.

LEMMA 7.34. *Let (X, d) be a metric space. If $A \subset X$ and c is a limit point of A, then for every $\varepsilon > 0$ the open ball $B_d(c, \varepsilon)$ intersects A in infinitely many points.*

Proof. Let c be a limit point of A. We establish the desired result by contradiction. Thus suppose that there is an open ball $B_d(c, \varepsilon^*)$ that intersects A in at most finitely many points, say $a_1, \dots, a_n$. There is at least one point in this list that is not equal to c since c is a limit point of A. Therefore

$$\mu = \min\{d(c, a_i) \mid i = 1, \dots, n \text{ and } a_i \neq c\}$$

is defined and positive. It follows that $B_d(c, \frac{\mu}{2})$ contains no points of A except possibly c. But this contradicts c being a limit point of A. Therefore for every $\varepsilon > 0$, the open ball $B_d(c, \varepsilon)$ intersects A in infinitely many points. ∎

The second lemma that we need is a Lebesgue Number Lemma for limit point compact metric spaces. We are in a somewhat peculiar situation here. We have Lemma 7.27, the Lebesgue Number Lemma for compact metric spaces, and compactness is equivalent to limit point compactness for metric spaces, so does that give us a Lebesgue Number Lemma for limit point compact metric spaces? The answer is no. We have not yet shown that compactness and limit point compactness are equivalent for metric spaces; in fact, that is what we are trying to prove here. Thus, we need to go through the work of proving a Lebesgue Number Lemma for limit point compact spaces, even though it is ultimately equivalent to the Lebesgue Number Lemma that we already proved.

LEMMA 7.35. *Let (X, d) be a limit point compact metric space, and let $\mathcal{O}$ be an open cover of X. Then there is a $\lambda > 0$ such that for every $x \in X$ there exists a $U \in \mathcal{O}$ satisfying $B_d(x, \lambda) \subset U$.*

Proof. Let (X, d) be a limit point compact metric space, and assume that $\mathcal{O}$ is an open cover of X. Since $\mathcal{O}$ covers X, it follows that for each $x \in X$ there exists a $U_x \in \mathcal{O}$ such that $x \in U_x$. Let $\varepsilon_x > 0$ be such that $B_d(x, \varepsilon_x) \subset U_x$.

To derive a contradiction, assume that no λ exists as specified in the statement of the lemma. Then for each $n \in \mathbb{Z}_+$, we can choose an $x_n \in X$ such that $B_d(x_n, \frac{1}{n})$ is contained in no set $U \in \mathcal{O}$.

There may be repetition among the points x_n, but we claim that the set $Y = \{x_n\}_{n \in \mathbb{Z}_+}$ contains infinitely many different points. We prove this claim by showing that there is no $p \in X$ such that $p = x_n$ for infinitely many n. To derive a contradiction, assume that there is such a point p. Then since $p = x_n$ for infinitely many n, we can choose $M \in \mathbb{Z}_+$ such that $p = x_M$ and $\frac{1}{M} < \varepsilon_p$. Now, by definition of ε_p there exists a $U_p \in \mathcal{O}$ such that $B_d(p, \varepsilon_p) \subset U_p$. Furthermore, $B_d(p, \frac{1}{M}) \subset B_d(p, \varepsilon_p)$ since $\frac{1}{M} < \varepsilon_p$. Therefore U_p is a set in $\mathcal{O}$ such that $B_d(p, \frac{1}{M}) \subset U_p$. But since $p = x_M$, this contradicts the definition of x_M. It follows that there is no $p \in X$ such that $p = x_n$ for infinitely many n, and therefore Y is an infinite set.

Since X is limit point compact, Y has a limit point y. Now $B_d(y, \varepsilon_y)$ is contained in some set $O_y \in \mathcal{O}$. By Lemma 7.34, since y is a limit point of Y, infinitely many of the x_n's lie in $B_d(y, \frac{\varepsilon_y}{2})$. Choose $N \in \mathbb{Z}_+$ such that $x_N \in B_d(y, \frac{\varepsilon_y}{2})$ and $\frac{1}{N} < \frac{\varepsilon_y}{2}$. It follows that $B_d(x_N, \frac{1}{N}) \subset B_d(y, \varepsilon_y) \subset O_y$, contradicting the definition of x_N. Thus there exists a λ as specified in the statement of the lemma. ∎

Now, we are ready to prove the desired theorem.

THEOREM 7.36. *If (X, d) is a metric space, then X is compact if and only if it is limit point compact.*

Proof. Having already established Theorem 7.33, we only need to prove that if X is a limit point compact metric space, then X is compact. Thus, assume that X is a limit point compact metric space with metric d, and let $\mathcal{O}$ be an open cover of X. We show that $\mathcal{O}$ has a finite subcover.

By Lemma 7.35 there exists a Lebesgue number λ for $\mathcal{O}$. Consider the cover of X given by

$$\mathcal{C} = \{B_d(x, \lambda) \mid x \in X\}.$$

We claim that $\mathcal{C}$ has a finite subcover. We use such a subcover of $\mathcal{C}$ to help us obtain a finite subcover of $\mathcal{O}$.

To prove that $\mathcal{C}$ has a finite subcover, we assume that it does not and derive a contradiction. Thus assume that no finite subcollection of $\mathcal{C}$ covers X. Pick $x_1 \in X$. Then $B_d(x_1, \lambda)$ does not cover X, and therefore we can pick $x_2 \in X - B_d(x_1, \lambda)$. Now, $\{B_d(x_1, \lambda), B_d(x_2, \lambda)\}$ does not cover X either, so we can pick $x_3 \in X - (B_d(x_1, \lambda) \cup B_d(x_2, \lambda))$. Continuing this process, we define a set $Y = \{x_n\}_{n \in \mathbb{Z}_+} \subset X$ that is such that $x_k \notin \bigcup_{j=1}^{k-1} B_d(x_j, \lambda)$ for all $k \in \mathbb{Z}_+$. (See Figure 7.17.)
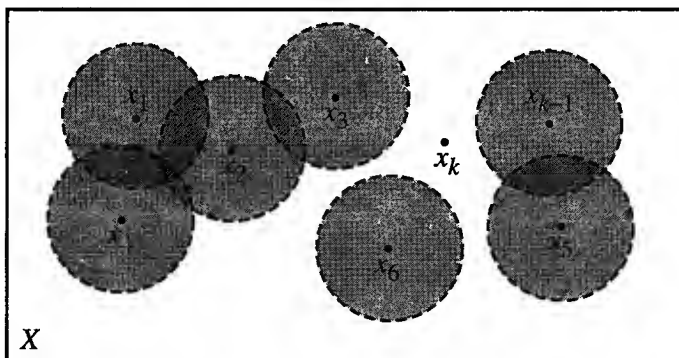

FIGURE 7.17: The point x_k does not lie in any of the balls $B_d(x_i, \lambda)$ for $i < k$.

The points in Y are distinct. Therefore Y is an infinite set. Let y be a limit point of Y. Then by Lemma 7.34, $B_d(y, \frac{\lambda}{2})$ contains infinitely many points from Y. Let x_p and x_q be two such points and assume that $p < q$. Since x_p and x_q both lie in $B_d(y, \frac{\lambda}{2})$, it follows that $d(x_p, x_q) < \lambda$. But this contradicts the fact that $x_q \notin \bigcup_{j=1}^{q-1} B_d(x_j, \lambda)$. Therefore we may infer that $\mathcal{C}$ has a finite subcover.

Let $\{B_1, \ldots, B_m\}$ be a subcover of $\mathcal{C}$. Each B_i is an open ball of radius λ. Since λ is a Lebesgue number for the cover $\mathcal{O}$, it follows that for each $i = 1, \ldots, m$, there exists an $O_i \in \mathcal{O}$ such that $B_i \subset O_i$. The collection $\{O_1, \ldots, O_m\}$ is then a finite subcover of $\mathcal{O}$. Therefore X is compact. ∎

Exercises for Section 7.4

7.33. Let $\mathbb{R}_d$ denote $\mathbb{R}$ with the discrete topology. Using the definition of limit point compactness directly, determine which of the subsets $A = \{(x, 2) \mid 1 \leq x \leq 2\}$ and $B = \{(3, y) \mid 1 \leq y \leq 2\}$, as previously shown in Figure 7.12, is a limit point compact subset of $\mathbb{R}_d \times \mathbb{R}$.

7.34. Let $Y = \{1, 2\}$ with the trivial topology and $\mathbb{Z}_+$ be the positive integers with the discrete topology. Prove that the product space $Y \times \mathbb{Z}_+$ is limit point compact but not compact.

7.35. Show that $[0, 1]$ is not limit point compact as a subspace of $\mathbb{R}$ with the lower limit topology.

7.36. Let X be a limit point compact space and A be a closed subset of X. Prove that A is limit point compact in the subspace topology.

7.5   One-Point Compactifications

As we have already seen, compact spaces and sets have a number of useful features. For example

- **(i)** Compact sets are closed and bounded in a metric space,

- **(ii)** Sequences have convergent subsequences in a compact subset of a metric space,

- **(iii)** Compact metric spaces are complete, and

- **(iv)** Continuous functions on compact spaces attain minimum and maximum values.

Furthermore, we have also encountered the useful properties possessed by a Hausdorff space. For example

- **(i)** Single-point sets are closed in a Hausdorff space, and

- **(ii)** Convergent sequences converge to a unique limit in a Hausdorff space.

Unfortunately, we do not always have the advantages afforded by a compact and Hausdorff space in the topological spaces we use. In this section, we introduce the property of local compactness and a construction, called the one-point compactification, that allows us to add a single point to a locally compact Hausdorff space X in order to obtain a compact Hausdorff space Y containing X as a subspace. In the case where $X = \mathbb{R}^3$, we will see that the one-point compactification yields the 3-sphere S^3.

DEFINITION 7.37. *A topological space X is **locally compact** if every $x \in X$ has a neighborhood that is contained in a compact subset of X.*

EXAMPLE 7.16. Every compact space is automatically locally compact since each $x \in X$ has X both as a neighborhood and as a compact set containing the neighborhood.

EXAMPLE 7.17. The real line $\mathbb{R}$ is locally compact since for each $x \in \mathbb{R}$ we have $x \in (x - 1, x + 1) \subset [x - 1, x + 1]$, and $[x - 1, x + 1]$ is compact.

EXAMPLE 7.18. The subspace $\mathbb{Q}$ of $\mathbb{R}$ in the standard topology is not locally compact. (See Exercise 7.37.)

DEFINITION 7.38. *Let X be a Hausdorff space. Set Y equal to the union of X and a single additional point, denoted ∞. (See Figure 7.18.) Define the open sets for a topology on $Y = X \cup \{\infty\}$ to be subsets of the following two types:*

    (i) *Open sets in X, and*

    (ii) *Sets of the form $Y - C$, where C is a compact subset of X.*

    *We call the resulting topological space Y the **one-point compactification** of X.*
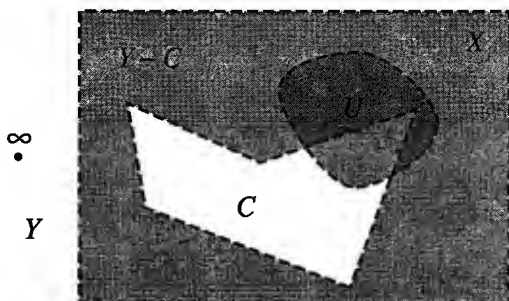
FIGURE 7.18: Open sets U and $Y - C$ in the one-point compactification $Y = X \cup \{\infty\}$.

Of course, we need to verify that the collection of open sets just described is a topology. We do that next.

THEOREM 7.39. *Let X be a Hausdorff space. The collection of subsets of $Y = X \cup \{\infty\}$ in the definition of the one-point compactification of X is a topology on Y.*

Proof. The empty set is open in Y since it is an open subset of X. The entire set Y itself is open in Y since it is the complement of $\varnothing$ in Y, and $\varnothing$ is a compact subset of X.

To prove that finite intersections of open sets in Y are open in Y, it is enough to check intersections of pairs of open sets U and V. The result for arbitrary finite intersections then follows by induction. Thus assume that U and V are open sets in Y. We need to check three separate cases. First, if both U and V are open sets in X, then $U \cap V$ is an open set in X, making it an open set in Y. Second, assume that $U = Y - C_1$ and $V = Y - C_2$, where C_1 and C_2 are compact subsets of X. Then $U \cap V = Y - (C_1 \cup C_2)$. Since finite unions of compact sets are also compact, $C_1 \cup C_2$ is a compact subset of X. It follows that $U \cap V = Y - C$ for a compact subset C of X, and therefore $U \cap V$ is open in Y in this case as well. Finally, assume that U is an open set in X and $V = Y - C$, where C is a compact subset of X. Then since ∞ is not in U, it follows that $U \cap V = U \cap (X - C)$. Now C is closed in X by Theorem 7.8 since it is a compact set in the Hausdorff space X. Therefore $X - C$ is open in X, implying that $U \cap (X - C)$ is open in X. Thus, $U \cap V$ is open in X, making it open in Y in this case, too. It follows that if U and V are arbitrary open sets in Y, then $U \cap V$ is also open in Y, as we wished to show.

Last, we prove that arbitrary unions of open sets are open. We can express such an arbitrary union in the form

$$(\bigcup_{\alpha \in A} U_\alpha) \cup (\bigcup_{\beta \in B}(Y - C_\beta)),$$

where each U_α is open in X and each C_β is a compact subset of X. The set $\bigcup_{\alpha \in A} U_\alpha$ is open in X; we denote this by U. Furthermore,

$$\bigcup_{\beta \in B}(Y - C_\beta) = Y - \bigcap_{\beta \in B} C_\beta,$$

and since each set C_β is a compact subset of the Hausdorff space X, Theorem 7.6 implies that $\bigcap_{\beta \in B} C_\beta$ is a compact subset of X. Letting $C = \bigcap_{\beta \in B} C_\beta$, we see that $\bigcup_{\beta \in B}(Y - C_\beta) = Y - C$ and C is a compact subset of X. Therefore we only need to verify that $U \cup (Y - C)$ is open in Y where U is open in X and C is a compact subset of X. Let $C' = X - U$, the complement of U in X; then C' is closed in X and therefore

$$\begin{aligned} U \cup (Y - C) &= (X - C') \cup (Y - C) \\ &= (Y - C') \cup (Y - C) \\ &= Y - (C' \cap C). \end{aligned}$$

Now C is a compact subset of the Hausdorff space X, so C is closed in X. Therefore $C' \cap C$ is closed in X, and since $C' \cap C$ is a subset of the compact set C, it follows that $C' \cap C$ is a compact subset of X. Thus

$Y - (C' \cap C)$ is an open set in Y, implying that $U \cup (Y - C)$ is an open set in Y. It follows that an arbitrary union of open sets in Y is an open set in Y. Thus the collection of subsets of Y described in the definition of the one-point compactification is a topology on Y.   ∎

Now, X is a subset of the one-point compactification $Y = X \cup \{\infty\}$. Therefore X inherits a subspace topology from Y. The next theorem indicates that this subspace topology is the same as the original topology, and therefore we can view X as a subspace of its one-point compactification.

THEOREM 7.40. *Let X be a Hausdorff space, and let $Y = X \cup \{\infty\}$ be its one-point compactification. Then the subspace topology that X inherits from Y is equal to the original topology on X.*

Proof. See Exercise 7.38.   ∎

Next, we justify the term "compactification."

THEOREM 7.41. *Let X be a Hausdorff space. Its one-point compactification $Y = X \cup \{\infty\}$ is compact.*

Proof. Let $\mathcal{O}$ be an open cover of Y. Define $\mathcal{O}_X$ to be the collection of subsets of X given by $\{V \cap X \mid V \in \mathcal{O}\}$. The sets in $\mathcal{O}_X$ are open in the subspace topology that X inherits from Y; therefore, by Theorem 7.40, they are open sets in X. It follows that $\mathcal{O}_X$ is an open cover of X.

Now, $\mathcal{O}$ is an open cover of Y, and therefore there exists a $U \in \mathcal{O}$ such that $\infty \in U$. It must be that $U = Y - C$ where C is a compact subset of X. The collection $\mathcal{O}_X$ covers C. Since C is a compact subset of X it follows that there is a finite collection $\{V_1 \cap X, \dots, V_n \cap X\} \subset \mathcal{O}_X$ that covers C. Therefore $\{O, V_1, \dots, V_n\}$ is a finite subcover of $\mathcal{O}$, implying that Y is compact.   ∎

Although we can construct the one-point compactification of any Hausdorff space, the result is not necessarily Hausdorff. For example, the one-point compactification of the set of rational numbers, $\mathbb{Q}$, as a subspace of $\mathbb{R}$, is not a Hausdorff space. (See Exercise 7.40.) On the other hand, we have the following theorem:

THEOREM 7.42. *Let X be a locally compact Hausdorff space. Then $Y = X \cup \{\infty\}$, the one-point compactification of X, is Hausdorff.*

Proof. To see that Y is Hausdorff, let x and y be points in Y. In the first case, assume both x and y are in X. Since X is Hausdorff, we can find disjoint open sets U and V in X that contain x and y, respectively. The sets U and V are also open sets in Y, and therefore there exist disjoint neighborhoods of x and y in Y. In the second case, suppose that $x = \infty$

and $y \in X$. By the local compactness of X, there is a compact set C in X that contains a neighborhood U of y. Now, $Y - C$ and U are open sets in Y, are disjoint, and contain x and y, respectively. Thus, in this case too, there exist disjoint neighborhoods of x and y in Y. It follows that Y is Hausdorff. ∎

EXAMPLE 7.19. The one-point compactification of the plane is homeomorphic to the sphere, S^2. In Example 4.16 we showed that the 2-sphere with the north pole, N, removed is homeomorphic to the plane via stereographic projection. The one-point compactification of the plane essentially fills the missing point, N, back in. To make this explicit, we define a function f from the sphere S^2 to the one-point compactification $\mathbb{R}^2 \cup \{\infty\}$ by taking each point in $S^2 - \{N\}$ to the corresponding point in the plane by stereographic projection and by taking N to ∞. (See Figure 7.19.) The resulting function f is a homeomorphism.
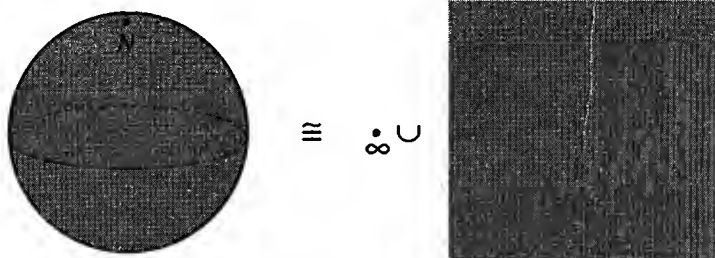
FIGURE 7.19: The sphere is homeomorphic to the one-point compactification of the plane.

When the plane is considered to be the complex plane $\mathbb{C}$, then the one-point compactification $\mathbb{C} \cup \{\infty\}$ is called the **Riemann sphere** or the **extended complex plane**. In this setting, the addition of the point ∞ not only plays a topological role, but also an algebraic role (for example, $\frac{1}{0} = \infty$), and the resulting space serves as a domain for many functions studied in complex analysis.

EXAMPLE 7.20. The 3-sphere, S^3, is the set of points in $\mathbb{R}^4$ that are at a distance 1 from the origin. It can be difficult to visualize the 3-sphere since we do not have natural experience visualizing 4-space. However, by an argument similar to the one that appears in the previous example, we can show that the 3-sphere S^3 is homeomorphic to the one-point compactification of $\mathbb{R}^3$.

When mathematicians study knot theory, they always assume that the embeddings of knots are in the compact space S^3 rather than the noncompact subspace $\mathbb{R}^3$. By adding just a single point to $\mathbb{R}^3$ they can take advantage of the benefits of working within a compact setting.

Exercises for Section 7.5

7.37. Show that $\mathbb{Q}$ is not locally compact in the subspace topology it inherits from $\mathbb{R}$ in the standard topology.

7.38. Prove Theorem 7.40: Let X be a Hausdorff space, and let $Y = X \cup \{\infty\}$ be its one-point compactification. Then the subspace topology that X inherits from Y is equal to the original topology on X.

7.39. Show that the one-point compactification of $(0, 1)$ is homeomorphic to the circle.

7.40. Show that the one-point compactification of $\mathbb{Q}$ is not Hausdorff.

7.41. (a) Describe and illustrate the result of taking the one-point compactification of the open annulus $S^1 \times (0, 1)$.

   (b) An **open Möbius band** is the space obtained from $[0, 1] \times (0, 1)$ by gluing the ends as we do with the usual Möbius band. Describe and illustrate the result of taking the one-point compactification of the open Möbius band. (Hint: The resulting space is one that we have previously encountered.)

7.42. Let X be Hausdorff and assume $Y = X \cup \{\infty\}$ is the one-point compactification of X.

   (a) Show that if X is not compact, then $\mathrm{Cl}(X) = Y$.

   (b) Show that if X is compact, then $\mathrm{Cl}(X) = X$, and Y is disconnected with $\{\infty\}$ being one of its components. (This shows that not much interesting happens when taking the one-point compactification of a space that is already compact.)

Dynamical Systems and Chaos

The area of mathematics known as dynamical systems has seen tremendous interest and growth over the past thirty years. A dynamical system is a particular type of function used to model time-varying processes. Examples of such processes appear in fluid mechanics, population growth, celestial mechanics, cardiac behavior, particle dynamics, and a multitude of situations where a physical system changes over time. In the analysis of a specific dynamical system, qualitative tools and techniques are often employed, and concepts from topology often underlie these qualitative methods.

Considered one of the leading figures in the development of topology in the late 1800s and early 1900s, Henri Poincaré (1854–1912) is also generally regarded as the principal founder of the field of dynamical systems. Poincaré studied the three-body problem, modeling the positions and velocities of three bodies in motion under each other's gravitational influence. Because general-solution formulas for the associated differential equations are difficult to obtain, he took the novel approach of qualitatively studying their structure within the space in which they are defined. Thus the field of dynamical systems was born.

In this chapter, we consider dynamical systems defined by repeated application of a given function that maps a space to itself. In the first two sections we examine basic properties of such systems. In Section 8.3, we provide a topological definition of a chaotic dynamical system and explore the relatively new idea that a system can be completely deterministic, with all of its behavior specified by initial conditions and simple rules for its evolution, and yet still exhibit the unpredictable behavior known as chaos. In Section 8.4, we present a simple population model that helped set off the chaos revolution. Finally, in Section 8.5, we prove that sensitive dependence on initial conditions (popularly referred to as "the butterfly effect") is a consequence of the topological definition of chaos.

8.1 Iterating Functions

We focus here on dynamical systems defined by repeated application of a function that maps a space to itself. Specifically, let X be a topological space and $f : X \to X$ be a function mapping X to itself. For every $n \in \mathbb{Z}_+$, define $f^n(x) = f \circ f \circ \ldots \circ f(x)$, the composition of n copies of the function f. The idea is that we start with x, then apply f to x, then apply f to $f(x)$, and continue this iterative process until we obtain $f^n(x)$.

DEFINITION 8.1. *The dynamical system defined by* $f : X \to X$ *is the family of functions* $\{f^n\}_{n \in \mathbb{Z}_+}$, *with each* f^n *mapping* X *to* X.

EXAMPLE 8.1. Let $f : \mathbb{R} \to \mathbb{R}$ be given by $f(x) = x/3$. Then the dynamical system defined by f is the family of functions given by $f^n(x) = x/3^n$.

In an application of a dynamical system defined by iterating a function $f : X \to X$, we think of $f(x)$ as describing the new state of the system one unit of time after it was at state x. For example, if we are modeling a bacteria population growing by the hour, we might have a function $f(x)$ representing the population size that results one hour after the population was x. Or, if we are modeling the position and velocity of a rocket, we might have a function $f(x, v)$ representing the position and velocity of the rocket one second after it had position and velocity (x, v).

Let us take a look at a few more examples of functions defining a dynamical system.

EXAMPLE 8.2. Consider the following four functions defined on $\mathbb{R}$:

(i) $f : \mathbb{R} \to \mathbb{R}$, defined by $f(x) = -2x$,

(ii) $g : \mathbb{R} \to \mathbb{R}$, defined by $g(x) = \frac{1}{2}x$,

(iii) $h : \mathbb{R} \to \mathbb{R}$, defined by $h(x) = -x$,

(iv) $k : \mathbb{R} \to \mathbb{R}$, defined by $k(x) = 0$.

When each of these functions is evaluated at $x = 0$, the result is 0. We say that 0 is a fixed point for the associated dynamical system. (We will subsequently define "fixed point" and other terms used in this example.)

First, consider the function f. If we take a particular point x_0, then $f^n(x_0) = (-2)^n x_0$. Therefore if $x_0 \neq 0$, then repeated iteration of f on x_0 results in values that move further and further from 0, bouncing back and forth between positive and negative values. The dynamics of f on $\mathbb{R}$ are qualitatively depicted in Figure 8.1 in what is called a **phase diagram** for the dynamical system.
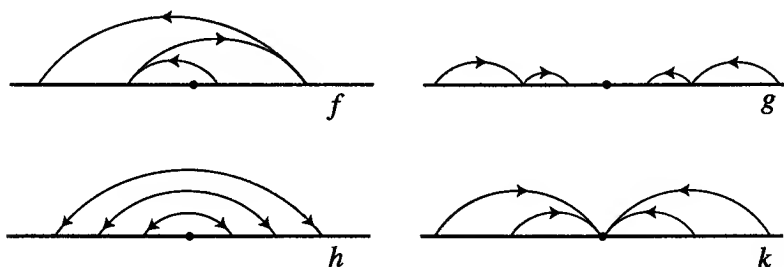
FIGURE 8.1: Phase diagrams for f, g, h, and k.

Next consider g. Here iteration of g on a nonzero value results in values that move progressively closer to 0, approaching 0 in the limit. In this case, 0 is referred to as an *asymptotically stable fixed point*.

With h we see a different dynamic picture. We have the fixed point at 0, but with any other x_0 the result of iterating h is an oscillation between the values $-x_0$ and x_0. Each nonzero value x_0 is called a *period-2 point* of the dynamical system.

Finally, consider k. Here the dynamics are simple. Every point, upon application of k, is sent immediately to the fixed point at 0. So 0 is a fixed point and every other point is referred to as an *eventual fixed point*.

DEFINITION 8.2. *Let $f : X \to X$, and assume $x \in X$.*

   (i) *The **orbit of x under** f is the sequence*

$$(x, f(x), f^2(x), \dots, f^n(x), \dots)$$

   and is denoted $O(x)$.

   (ii) *We say that x is a **fixed point of** f if $f(x) = x$. So the orbit of a fixed point x is a constant sequence at the point x.*

   (iii) *We say that x is an **eventual fixed point of** f if x is not a fixed point of f but $f^n(x)$ is a fixed point for some $n \in \mathbb{Z}_+$.*

   (iv) *Assume $m \in \mathbb{Z}_+$. We say that x is a **periodic point** or a **period-m point** if $f^m(x) = x$ and $f^j(x) \neq x$ for $j = 1, \dots, m-1$. Under these circumstances the orbit of x is called a **periodic orbit** or a **period-m orbit**. Also, we say that m is the **period** of the periodic point or the periodic orbit.*

   (v) *We say that x is an **eventual periodic point** if x is not a periodic point but $f^n(x)$ is a periodic point for some $n \in \mathbb{Z}_+$.*

If x is a period-m point for a function $f : X \to X$, then the points $x, f(x), \dots, f_{m-1}(x)$ are all period-m points and are all distinct. (See Exercise 8.3.) Thus, iterating f on x, we continually cycle through these m periodic points.

A fixed point is a period-1 point, so we include the fixed points within the set of periodic points of a function.

EXAMPLE 8.3. Here we consider two simple examples of savings accounts. First, suppose that we deposit money in a savings account that earns 5% interest, compounded annually. After the initial deposit, we do not make any further deposits to the account nor do we make any withdrawals from it. We simply let the amount in the account accrue the earned interest. The function $f : [0, \infty) \to [0, \infty)$, given by $f(x) = 1.05x$, defines a dynamical system

that models the amount in the account as it changes year by year. The dynamics of f are straightforward: there is a fixed point at 0, and every other point has an orbit that increases away from 0 upon successive iteration of f. (See Figure 8.2.)
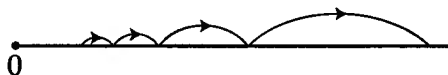

FIGURE 8.2: The dynamics of f.

Now assume that after the interest is applied each year, we withdraw either $1,000 from the account (if there is at least that much in the account) or the balance of the account (if it is less than $1,000). In this case the function $g : [0, \infty) \rightarrow [0, \infty)$, given by

$$g(x) = \begin{cases} 1.05x - 1000 & \text{if } 1.05x \geq 1000, \\ 0 & \text{if } 1.05x \leq 1000, \end{cases}$$

defines a dynamical system modeling how the amount in the account changes. Here too, 0 is a fixed point. There is another fixed point at $x = 20,000$ that we find by solving $g(x) = x$. We can also find the fixed point at 20,000 by reasoning that the amount in the account will be fixed when it is such that the interest of 5% provides exactly the $1,000 needed for the annual withdrawal. Since 5% of 20,000 is 1,000, it follows that the fixed point occurs at $x = 20,000$. For values of x greater than 20,000, the interest on x provides more than the amount needed for the $1,000 withdrawal, so the amount in the account will grow without bound under successive iteration of g. If $x < 20,000$, then eventually the amount in the account will equal 0. So nonzero values of x that are less than 20,000 are eventual fixed points of g. We illustrate the dynamics of g in Figure 8.3.
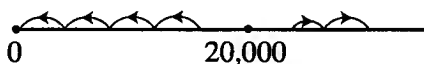
FIGURE 8.3: The dynamics of g.

As we saw in Example 8.3, if we have a dynamical system defined by iterating a function f, then we can find fixed points by solving the equation $f(x) = x$. Similarly, to find a period-m point, we solve the equation $f^m(x) = x$ and take the solutions that are not solutions of $f^j(x) = x$ for $j = 1, \ldots, m - 1$.

EXAMPLE 8.4. Imagine a batch of raisin-bread dough lying across the interval $[0, 1]$. Suppose that we uniformly stretch the dough to twice its length and then fold the dough over, pressing it together so that it again lies across the interval $[0, 1]$. (See Figure 8.4.)
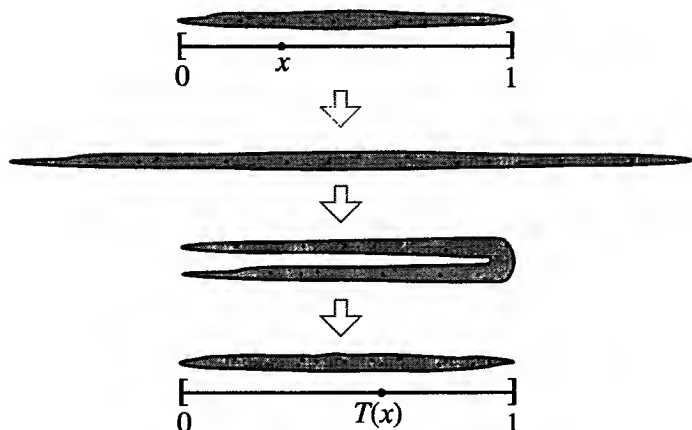
FIGURE 8.4: Stretching, folding, and pressing the raisin-bread dough.

Let $T : [0, 1] \rightarrow [0, 1]$ be defined by setting $T(x)$ equal to the new position of a raisin that was originally at x, after the stretching, folding, and pressing of the dough. More precisely, T is defined by

$$T(x) = \begin{cases} 2x & \text{if } x \in [0, \frac{1}{2}], \\ 2 - 2x & \text{if } x \in [\frac{1}{2}, 1]. \end{cases}$$

The graph of T is shown in Figure 8.5. For obvious reasons, T is referred to as the **tent function**. We also show the graph of the line $y = x$ in Figure 8.5. Fixed points of T occur where $T(x) = x$ and therefore at values of x where the graph of T intersects the line $y = x$. We can see that T has two fixed points; they are located at 0 and 2/3.
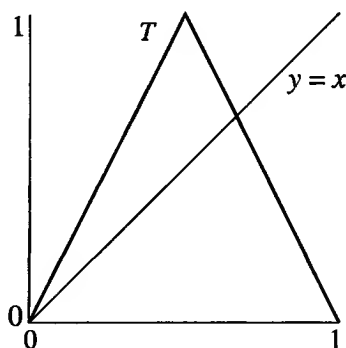
FIGURE 8.5: The tent function.

It is straightforward to see that 1 and 1/2 are eventual fixed points of T, but these are not the only ones—there are many, many more, and we show how to identify all of them in Section 8.3.

Now consider T^2. The graph is shown in Figure 8.6, along with the graph of $y = x$. We see that there are four solutions to $T^2(x) = x$; two of them are the fixed points 0 and 2/3, and the other two are at 2/5 and 4/5. The latter two are period-2 points and together they form a periodic orbit.

Here we have only touched on the very complicated dynamics of the tent function. We explore T further in Section 8.3, where we show that T is chaotic.
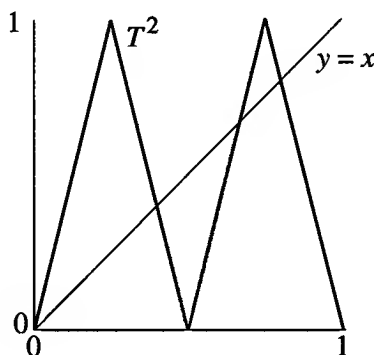
FIGURE 8.6: The function T^2.

The following definition establishes a notion of equivalence for dynamical systems defined by function iteration:

DEFINITION 8.3. *The functions $f : X \rightarrow X$ and $g : Y \rightarrow Y$ (and the dynamical systems defined by them) are said to be **topologically conjugate** if there exists a homeomorphism $h : X \rightarrow Y$ such that $g \circ h = h \circ f$. The function h is called a **topological conjugacy** between f and g.*

We illustrate the topological conjugacy condition $g \circ h = h \circ f$ in Figure 8.7. The idea is that both routes from the upper-left X to the lower-right Y—across the top, then down the right side, and down the left side, then across the bottom—give the same result. We say that the diagram commutes. Essentially, h is mapping the function f to the function g.
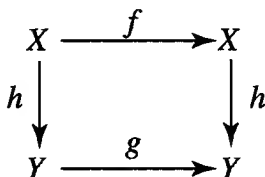
FIGURE 8.7: Topological conjugacy requires that $g \circ h = h \circ f$.

EXAMPLE 8.5. The dynamics of the functions $f(x) = 2x$ and $g(x) = 3x$ appear qualitatively the same. In both cases there is a fixed point at 0, and all other orbits stay either on the positive or negative side of 0 and move outward

from 0. In fact, these two functions are topologically conjugate. The function $h : \mathbb{R} \to \mathbb{R}$, defined by $h(x) = x^{\log_2(3)}$, is a homeomorphism that satisfies $g \circ h = h \circ f$.

A topological conjugacy between two functions f and g naturally maps orbits of f to orbits of g, as the following theorem indicates:

THEOREM 8.4. *Let h be a topological conjugacy between $f : X \to X$ and $g : Y \to Y$. For each $x \in X$ and $n \in \mathbb{Z}_+$, we have $h(f^n(x)) = g^n(h(x))$, and consequently h maps the orbit of x under f to the orbit of $h(x)$ under g.*

> **Proof.** We prove this by induction on n. The $n = 1$ case holds by the definition of topological conjugacy. Assume that the result holds for $n - 1$. Then,
>
> $$\begin{aligned} h(f^n(x)) &= h(f^{n-1}(f(x))) \\ &= g^{n-1}(h(f(x))) \\ &= g^{n-1}(g(h(x))) \\ &= g^n(h(x)), \end{aligned}$$
>
> where the second equality holds by the inductive hypothesis. Thus if the result holds for $n - 1$, then it holds for n. Therefore, by induction, $h(f^n(x)) = g^n(h(x))$ for all $n \in \mathbb{Z}_+$. ∎

The following corollary is an immediate consequence of Theorem 8.4:

COROLLARY 8.5. *Let h be a topological conjugacy between $f : X \to X$ and $g : Y \to Y$, and assume that $x \in X$. Then the following implications hold:*

(i) If x is a fixed point of f, then $h(x)$ is a fixed point of g.

(ii) If x is a period-m point of f, then $h(x)$ is a period-m point of g.

(iii) If x is an eventual fixed point of f, then $h(x)$ is an eventual fixed point of g.

(iv) If x is an eventual periodic point of f, then $h(x)$ is an eventual periodic point of g.

> **Proof.** See Exercise 8.9. ∎

The corollary implies that important dynamic features of f are mirrored in functions that are topologically conjugate to f. We will encounter similar results throughout the chapter, and thereby see that topologically conjugate functions have equivalent dynamic behavior under iteration.

Exercises for Section 8.1

8.1. For each of the following functions $f : \mathbb{R} \to \mathbb{R}$, find all fixed points and periodic points and sketch a phase diagram for the dynamical system:

(a) $f(x) = x^3$

(b) $f(x) = -x^3$

(c) $f(x) = -x^{1/3}$

(d) $f(x) = x - x^2$

(e) $f(x) = \frac{4}{\pi} \tan^{-1}(x)$

(f) $f(x) = 1 - x^2$

(g) $f(x) = \frac{1}{2} \sin(x)$

(h) $f(x) = \frac{\pi}{2} \sin(x)$

(i) $f(x) = e^x$

(j) $f(x) = 2(x - x^2)$

(k) $f(x) = x + \sin(x)$

8.2. (a) Consider the linear systems $L_a : \mathbb{R} \to \mathbb{R}$ defined by $L_a(x) = ax$ where $a \in \mathbb{R}$. Through a collection of phase diagrams, classify the different dynamic behaviors seen in L_a, as a ranges over the real numbers.

(b) Show that if L_a and L_b have the same dynamic behavior (as identified in part (a)), then L_a and L_b are topologically conjugate.

8.3. Show that if x is a period-m point for a function $f : X \to X$, then the points $x, f(x), \ldots, f^{m-1}(x)$ are all distinct and are all period-m points.

8.4. Find as many eventual fixed points as you can for the tent function T. (In Section 8.3 we show how to identify all of them.)

8.5. Find a period-3 point for the tent function T.

8.6. Consider the second savings-account model in Example 8.3:

$$g(x) = \begin{cases} 1.05x - 1000 & \text{if } 1.05x \geq 1000, \\ 0 & \text{if } 1.05x \leq 1000. \end{cases}$$

(a) Determine $g^{-1}(0)$. That is, determine the interval of amounts that result in a balance of 0 the following year.

(b) Determine the interval of amounts that result in a nonzero balance after $n - 1$ years but a balance of 0 after n years.

8.7. In this exercise we ask you to explore, by explicit computation, orbits associated with the functions $f_\alpha : [0, 1] \to [0, 1]$, defined by $f_\alpha(x) = \alpha x (1 - x)$. We examine this family of functions further in Section 8.4.

(a) For $f_{0.6} : [0, 1] \to [0, 1]$, $f_{0.6}(x) = 0.6x(1 - x)$, explore and compare the orbits determined by the initial values 0, 0.1, and 0.7.

(b) For $f_{1.6} : [0, 1] \to [0, 1]$, $f_{1.6}(x) = 1.6x(1 - x)$, explore and compare the orbits determined by the initial values 0, 0.1, and 0.7.

(c) For $f_{3.2} : [0, 1] \to [0, 1]$, $f_{3.2}(x) = 3.2x(1 - x)$, explore and compare the orbits determined by the initial values 0.25 and 0.75.

(d) For $f_4 : [0, 1] \to [0, 1]$, $f_4(x) = 4x(1 - x)$, explore and compare the orbits determined by the initial values 0.261 and 0.262.

8.8. Find the fixed points of $g_a : \mathbb{R} \to \mathbb{R}$, $g_a(x) = a(x + x^2)$, for $a > 0$. On an x versus a coordinate system, plot the fixed points, demonstrating how their location changes as the parameter a changes.

8.9. **Prove Corollary 8.5:** Let h be a topological conjugacy between $f : X \to X$ and $g : Y \to Y$, and assume that $x \in X$.

(a) If x is a fixed point of f, then $f(x)$ is a fixed point of g.

(b) If x is a period-m point of f, then $f(x)$ is a period-m point of g.

(c) If x is an eventual fixed point of f, then $f(x)$ is an eventual fixed point of g.

(d) If x is an eventual periodic point of f, then $f(x)$ is an eventual periodic point of g.

8.10. Define the function $g : [0, 1] \to [0, 1]$ by

$$
g(x) = \begin{cases} 3x & \text{for } 0 \le x \le \frac{1}{3}, \\ 2 - 3x & \text{for } \frac{1}{3} \le x \le \frac{2}{3}, \\ 3x - 2 & \text{for } \frac{2}{3} \le x \le 1. \end{cases}
$$

Show that g is not topologically conjugate to the tent function.

8.11. Let $f : X \to X$ be continuous. Show that if $x \in X$ and $y = \lim_{n \to \infty} f^n(x)$, then y is a fixed point of f. That is, if the orbit of a point converges, then it converges to a fixed point. (Hint: Use Theorem 4.7.)

8.12. Let $f : \mathbb{R} \to \mathbb{R}$ be continuous, and assume that I and J are disjoint closed and bounded intervals such that $f(I) \subset J$ and $f(J) \subset I$. Prove that there is a period-2 point of f in I.

8.13. **(a)** Let $f : \mathbb{R} \to \mathbb{R}$ be a homeomorphism. Prove that f has no periodic points with period greater than 2.

(b) Show that for every $n \in \mathbb{Z}_+$ there exists a homeomorphism $f : \mathbb{R}^2 \to \mathbb{R}^2$ with a period-n point.

8.14. Assume that $f : X \to X$ is injective.
(a) Prove that $f^n : X \to X$ is injective for each $n \in \mathbb{Z}_+$.
(b) Prove that f has no eventual periodic points.

8.15. Consider $f(x) = \frac{1}{1-x}$. Let X be the maximal subspace of $\mathbb{R}$ on which f defines a dynamical system. Show that X has three components, and prove that each $x \in X$ is a period-3 point with an orbit having one point in each component of X.

8.2   Stability

In many physical settings and mathematical models, we observe behavior that converges toward fixed or periodic states. For example, a cake removed from an oven has a temperature that cools toward the fixed room temperature, or a child on a swing attains a steady oscillation as he or she pumps the swing to balance off the effect of damping. In mathematics, we see, for instance, that for $f(x) = \frac{1}{2}x$ every point has an orbit that converges to the fixed point at 0, and for $g(x) = -x^{1/3}$ every point except 0 has an orbit that converges to the periodic orbit $\{-1, 1\}$.

In this section we capture these behaviors through what is known as the asymptotic stability of a fixed point or periodic point of a dynamical system. We begin with the definitions associated with stability.

DEFINITION 8.6. *Given* $f : X \to X$, *assume that* x^* *is a fixed point of* f.

    (i) We say that x^* *is* **stable** *if for every open set* U *containing* x^* *there is an open set* V *containing* x^* *such that for every* $x \in V$, *the orbit of* x *lies in* U.

    (ii) We say that x^* *is* **asymptotically stable** *if* x^* *is stable and if there exists an open set* U *containing* x^* *such that*

$$\lim_{n \to \infty} f^n(x) = x^*$$

    for every $x \in U$.

    (iii) We say that x^* *is* **neutrally stable** *if* x^* *is stable but not asymptotically stable.*

    (iv) We say that x^* *is* **unstable** *if* x^* *is not stable.*

DEFINITION 8.7. *Given* $f : X \to X$, *assume that* x^* *is a period-m point of* f. *We say that* x^* *is a* **stable periodic point,** *or has a* **stable periodic orbit,** *if* x^* *is stable as a fixed point of* f^m. *We similarly define* **asymptotically stable,** **neutrally stable,** *and* **unstable** *for periodic points and periodic orbits.*

Implicit in the definition of stable periodic points and periodic orbits is the fact that if x^* is a stable period-m point of f, then so are

$$f(x^*), f^2(x^*), \ldots, f^{m-1}(x^*).$$

The same holds for the definitions of asymptotically stable, neutrally stable, and unstable. These facts are not automatic, but must be shown to be true for the definitions to make sense. (See Exercise 8.17.)

EXAMPLE 8.6. We consider again the four functions from Example 8.2:

    (i) $f : \mathbb{R} \to \mathbb{R}$, defined by $f(x) = -2x$,

    (ii) $g : \mathbb{R} \to \mathbb{R}$, defined by $g(x) = \frac{1}{2}x$,

    (iii) $h : \mathbb{R} \to \mathbb{R}$, defined by $h(x) = -x$,

    (iv) $k : \mathbb{R} \to \mathbb{R}$, defined by $k(x) = 0$.

There is a fixed point at 0 for each of these functions. For f, note that if we take the open interval $U = (-1, 1)$ containing 0, then, except for the orbit at 0, every orbit that begins in U eventually leaves U. Therefore 0 is an unstable fixed point of f. Next consider the fixed point at 0 for g. Given any open set U containing 0, let V be an open interval $(-\varepsilon, \varepsilon)$ contained in U. Every orbit beginning in V stays in V and approaches 0 in the limit. Therefore 0 is an asymptotically stable fixed point of g. For k, the fixed point at 0 is also asymptotically stable. For h, the fixed point at 0 is neutrally stable and every point other than 0 is a neutrally stable period-2 point.

EXAMPLE 8.7. Consider again the savings account models in Example 8.3. For the case where the account accumulates interest, and there are no deposits or withdrawals, we have the dynamical system defined by $f(x) = 1.05x$. For this f, we have an unstable fixed point at 0. If we start with any nonzero amount of money, no matter how small, the amount in the account grows away from 0.

In the second savings account model, where we have an annual \$1,000 withdrawal after the application of interest, the dynamical system is defined by

$$g(x) = \begin{cases} 1.05x - 1000 & \text{if } 1.05x \geq 1000, \\ 0 & \text{if } 1.05x \leq 1000. \end{cases}$$

In this case 0 is an asymptotically stable fixed point. If we start the account with amounts close to 0, then we end up at 0 in the long run. (in fact, for small enough initial amounts, we go to 0 immediately.) On the other hand, we have an unstable fixed point at 20,000. Initial amounts that are greater than 20,000 increase without bound under this dynamical system, and amounts less than 20,000 decrease to 0.

As already indicated, stable fixed points and periodic points are what we often observe or approach in the long run in both the mathematics of a dynamical system and the behavior of a physical system that a dynamical system models. We see this demonstrated in Examples 8.8 and 8.10.

EXAMPLE 8.8. Let $f : [0, 1] \to [0, 1]$ be defined by $f(x) = 2x(1 - x)$. If we take a value $w \in (0, 1)$ and use a calculator to compute $f(w)$ and successive iterations of f on w, we see that eventually the calculator returns 0.5 every time. Through this process we have discovered what appears to be a stable fixed point of f. In fact, 0.5 is an asymptotically stable fixed point, and this system drives all values in $(0, 1)$ toward it. Therefore being near 0.5, or fixed at 0.5, is expected long-term behavior in this system.

Of course, we could find the fixed point 0.5 for f by solving $2x(1 - x) = x$. In doing so, we obtain $x = 0$ or 0.5. So there is also a fixed point at 0. The fixed point at 0 is unstable, and therefore the process of successive iteration of f does not reveal this fixed point unless we choose it (or 1) at the start. Being fixed at 0 is not expected long-term behavior in this system.

Even though the calculator brings us to the value 0.5 after finitely many iterations of f on $w \in (0, 1)$, the point w is actually not an eventual fixed point of f. The reason that we eventually obtain exactly 0.5 is that the calculator rounds off its results along the way, and at some point it rounds off to exactly 0.5, where it is then fixed. So if $w \in (0, 1)$ and $w \neq 0.5$, then $\lim_{n \to \infty} f^n(w) = 0.5$, but $f^n(w) \neq 0.5$ for each n.

A convenient method for viewing the behavior of a dynamical system defined on a domain in $\mathbb{R}$ is what is known as a **web diagram**. (See Figure 8.8.)

We illustrate this method with the function $f : [0, 1] \to [0, 1]$ from the previous example. To begin, we draw the graphs of $f(x) = 2x(1-x)$ and $y = x$. Then we take an initial value $a \in [0, 1]$ and start with the point (a, a) on the line $y = x$. We go vertically to the point $(a, f(a))$ on the graph of $f(x)$ and then horizontally to the point $(f(a), f(a))$ on the line $y = x$. Repeating these steps brings us to the point $(f^2(a), f^2(a))$ on the line $y = x$. By further repeating this process, we obtain the orbit of a, depicted on the line $y = x$.
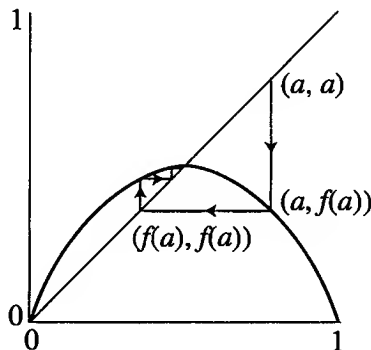
FIGURE 8.8: A web diagram for $f(x) = 2x(1-x)$.

The following theorem classifies all of the possible fixed-point and stability properties of the linear functions $f(x) = mx + b$:

THEOREM 8.8. *Let $f : \mathbb{R} \to \mathbb{R}$ be the linear function $f(x) = mx + b$.*

    (i) If $m \neq 1$, then f has a unique fixed point, and that fixed point is neutrally stable if $m = -1$, asymptotically stable if $|m| < 1$, and unstable if $|m| > 1$.

    (ii) If $m = 1$, then f has no fixed points if $b \neq 0$, and every $x \in \mathbb{R}$ is a neutrally stable fixed point if $b = 0$.

Proof. See Exercise 8.19. ∎

In Figure 8.9 we use web diagrams to illustrate some of the stability possibilities asserted in Theorem 8.8.

Now, if a function is differentiable at a fixed point, then the following theorem indicates that the stability of the fixed point can be determined by the value of the derivative at the fixed point, as long as that value is not -1 or 1.

THEOREM 8.9. *Let X be a subset of $\mathbb{R}$. Assume that x_0 is a fixed point of $f : X \to X$ and that f is differentiable at x_0. Then x_0 is asymptotically stable if $|f'(x_0)| < 1$ and x_0 is unstable if $|f'(x_0)| > 1$.*

We do not prove Theorem 8.9. Proofs can be found in standard introductory texts on dynamical systems, such as [DevR]. The idea is that as long as $|f'(x_0)| \neq 1$, the stability of x_0 as a fixed point of f is deter-
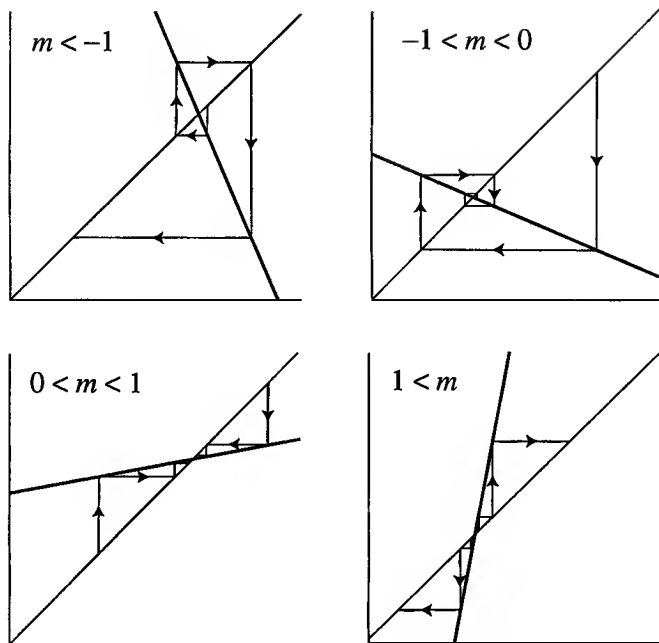
FIGURE 8.9: Stability properties of linear functions $f(x) = mx + b$.

mined by the stability of x_0 as a fixed point of the linear approximation to f based at x_0. (See Figure 8.10.) The linear approximation is the function $L(x) = f'(x_0)(x - x_0) + f(x_0)$.

In the transition cases that are not addressed by Theorem 8.9, where $|f'(x_0)| = 1$, it is possible that x_0 could be neutrally stable, asymptotically stable, or unstable. (See Exercise 8.20.)
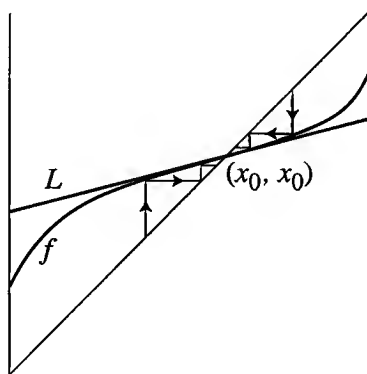
FIGURE 8.10: The stability of the fixed point x_0 is the same for both f and L.

EXAMPLE 8.9. Let us revisit some of the examples we have previously considered to see how Theorems 8.8 and 8.9 verify our observations regarding stability.

For the savings-account model without withdrawal in Example 8.3, given by $f(x) = 1.05x$, we have a fixed point at 0. The slope of f at 0 is 1.05, which is greater than 1, and therefore the fixed point is unstable.

For the savings-account model including withdrawal in Example 8.3, given by

$$g(x) = \begin{cases} 1.05x - 1000 & \text{if } 1.05x \geq 1000, \\ 0 & \text{if } 1.05x \leq 1000, \end{cases}$$

we have fixed points at 0 and 20,000. The slope of g at 20,000 is 1.05, implying that 20,000 is an unstable fixed point. The slope of g at 0 is 0, implying that 0 is an asymptotically stable fixed point.

For the function $f(x) = 2x(1 - x)$ in Example 8.8, there are fixed points at 0 and 1/2. Furthermore, $f'(0) = 2$, so 0 is an unstable fixed point, and $f'(1/2) = 0$, so 1/2 is a stable fixed point.

For the tent function T in Example 8.4 we have fixed points at 0 and 2/3. The slope of T at the former is 2 and at the latter is -2. It follows that both are unstable fixed points. There are also period-2 points at 2/5 and 4/5. We can see from the graph of T^2 in Figure 8.6 that the slope of T^2 at each is -4. Therefore these periodic points are unstable. Thus, in the tent function we do not expect to see the system settle down in the long run toward either of the fixed points at 0 and 2/3, nor do we expect it to settle down toward the periodic orbit $\{2/5, 4/5\}$. In fact, as we will see in Section 8.3, there are periodic points of T densely distributed over $[0, 1]$, but they are all unstable, so the system never settles down to them.

It is not a coincidence that the derivative of T^2 is the same at each of the points in the periodic orbit $\{2/5, 4/5\}$. In fact, given $f : X \to X$ where $X \subset \mathbb{R}$, if $x_1, \ldots, x_m$ form a period-m orbit of f, and if f is differentiable, then the derivative of f^m at each of the points $x_1, \ldots, x_m$ is equal to $f'(x_1)f'(x_2)\ldots f'(x_m)$, the product of the values obtained by taking the derivative of f at each point in the orbit. (See Exercise 8.21.)

EXAMPLE 8.10. In this example, we consider a mass that hangs on a spring attached to a ceiling. We assume that there is an equilibrium position at which the mass rests when there is no motion. The equilibrium position serves as the origin for determining the displacement of the mass. Further, we assume that when the mass is in motion, it oscillates up and down past the equilibrium position and rises no higher than a distance M from the equilibrium position. We model the motion of the mass by a function $f : [0, M] \to [0, M]$, which is defined so that $f(x)$ is the maximum height of the mass on a particular oscillation if the maximum height of the mass was x on the previous oscillation. (See Figure 8.11.)

In damped motion, for example that caused by air resistance, we would have $f(0) = 0$ and $f(x) < x$ for all other x, as shown on the left in Figure 8.12. Thus we would have a stable fixed point at 0, and in the long run we would expect to see the spring–mass system settle toward equilibrium.
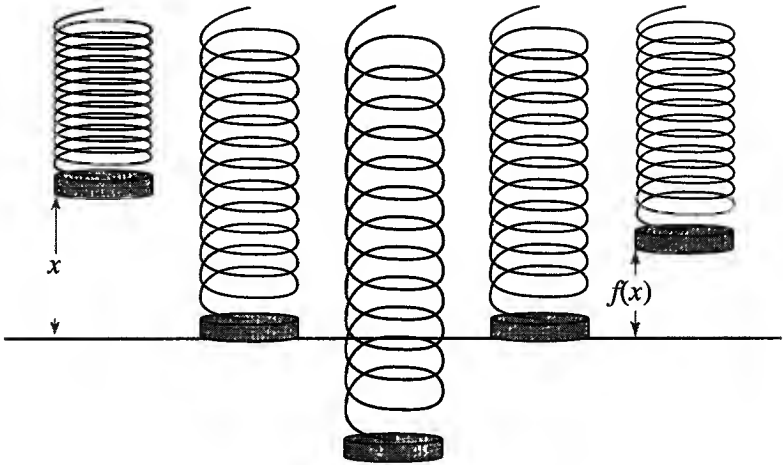
FIGURE 8.11: The spring and mass in oscillation.

On the other hand, assume that we are exciting the motion, for example by giving the spring a downward push every time it passes the equilibrium position in the downward direction. In this case we would have $f(0) = 0$ and $f(x) > x$ for all other x, as shown on the right in Figure 8.12. Here we would have an unstable fixed point at 0. Once this system was in motion it would stay in motion with successively greater and greater oscillations.
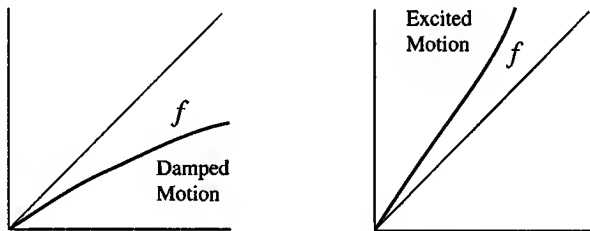

FIGURE 8.12: The graphs for damped motion and excited motion.

Let us see what happens when we combine these two scenarios, having small-scale excitation and large-scale damping. Assume we provide a push on the mass when the maximum points of the oscillations are near 0, and we let damping take hold when the maximum points are large. We assume our function f is such that $f(0) = 0$, $f(x) > x$ for x near 0, $f(x) < x$ for large x, and f is continuous as it transitions between the small-scale excitation and large-scale damping. (See Figure 8.13.) The Intermediate Value Theorem guarantees that we have a point $x_0 > 0$ with $f(x_0) = x_0$; that is, we have a fixed point. Note that the fixed point does not represent a situation where the mass is fixed, but instead represents a maximum height above equilibrium to which the mass returns at each oscillation. We regard this as a sustained oscillation—one that neither dampens nor grows. Thus, small-scale excitation with large-scale damping yields a sustained oscillation.

If, furthermore, the graph of f is as shown in Figure 8.13, then $0 < f'(x_0) < 1$ and x_0 is stable. Therefore the sustained oscillation is a behavior toward which we would expect this system to settle in the long run.
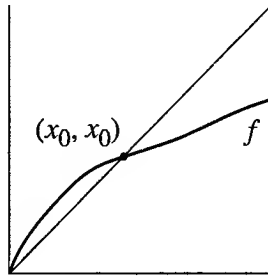
FIGURE 8.13: Combining small-scale excitation and large-scale damping.

Sustained oscillations, as seen in Example 8.10, are common in many physical systems. As you blow on the reed of a woodwind instrument, you are producing a small-scale excitation that is opposed by the restoring force tending to keep the reed flat. The oscillation that results from the balance between the excitation and the restoring force produces the sound that becomes music in the instrument. The van der Pol oscillator is a classic differential-equation model for a sustained oscillation. It was first introduced by Balthazar van der Pol (1889–1959) as a model for cardiac oscillations, but he also examined it more generally as a model for electronic circuits. The differential equation reflects both a small-scale excitation and a large-scale damping, and consequently its solutions tend toward a sustained oscillation. (See [Str].)

Exercises for Section 8.2

8.16. For each of the following functions determine the fixed and periodic points and for each such point use Theorem 8.9 to determine its stability:

(a) $f(x) = x^3$

(b) $f(x) = -x^3$

(c) $f(x) = -x^{1/3}$

(d) $f(x) = \frac{4}{\pi} \tan^{-1}(x)$

(e) $f(x) = 1 - x^2$

(f) $f(x) = \frac{1}{2} \sin(x)$

(g) $f(x) = \frac{\pi}{2} \sin(x)$

(h) $f(x) = 2(x - x^2)$

(i) $f(x) = x + \sin(x)$

8.17. (a) Prove that if x^* is a stable period-m point of a function f, then so are $f(x^*), f^2(x^*), \ldots, f^{m-1}(x^*)$.

(b) Prove that if x^* is an asymptotically stable period-m point of a function f, then so are $f(x^*), f^2(x^*), \ldots, f^{m-1}(x^*)$.

(c) Prove that if x^* is a neutrally stable period-m point of a function f, then so are $f(x^*), f^2(x^*), \ldots, f^{m-1}(x^*)$.

(d) Prove that if x^* is an unstable period-m point of a function f, then so are $f(x^*), f^2(x^*), \ldots, f^{m-1}(x^*)$.

8.18. Draw web diagrams for the dynamical systems in Examples 8.3, 8.4, and 8.8, illustrating the stability properties discussed in Example 8.9.

8.19. Prove Theorem 8.8: Let $f : \mathbb{R} \to \mathbb{R}$ be the linear function $f(x) = mx + b$.
   (a) If $m \neq 1$, then f has a unique fixed point, and that fixed point is neutrally stable if $m = -1$, asymptotically stable if $|m| < 1$, and unstable if $|m| > 1$.
   (b) If $m = 1$, then f has no fixed points if $b \neq 0$, and every $x \in \mathbb{R}$ is a neutrally stable fixed point if $b = 0$.

8.20. In the transition cases that are not addressed by Theorem 8.9, where $|f'(x_0)| = 1$ at a fixed point x_0 of f, it is possible that x_0 could be neutrally stable, asymptotically stable, or unstable.
   (a) Find an example of a function f with a fixed point x_0 where $|f'(x_0)| = 1$ and x_0 is neutrally stable.
   (b) Find an example of a function f with a fixed point x_0 where $|f'(x_0)| = 1$ and x_0 is asymptotically stable.
   (c) Find an example of a function f with a fixed point x_0 where $|f'(x_0)| = 1$ and x_0 is unstable.

8.21. Assume $X \subset \mathbb{R}$ and $f : X \to X$ is differentiable. Show that if $x_1, \ldots, x_m$ constitute a period-m orbit of f, then for each i,

$$(f^m)'(x_i) = f'(x_1) f'(x_2) \ldots f'(x_m).$$

8.22. Newton's method is an iterative process for approximating the zeros of a function $f : D \to \mathbb{R}$ with $D \subset \mathbb{R}$. We assume that f is twice differentiable. For such an f define $g(x) = x - \frac{f(x)}{f'(x)}$. The idea behind Newton's method is to make an initial guess x_0 of a zero of f, and then iterate g on x_0 to obtain a sequence of points, $x_n = g^n(x_0)$, that (we hope) converges to a zero of f.
   (a) Show that if x is a fixed point of g then x is a zero of f.
   (b) Show that every fixed point of g is asymptotically stable.

If x is such that $f'(x) = 0$ then g is not defined at x. Such an x could be a zero of f. If z is a zero of f for which $f'(z) \neq 0$, then the results of parts (a) and (b) imply that if we make an initial guess x_0 sufficiently close to z, then the orbit of x_0 under g converges to z.

Now, consider the situation where we have two fixed points z_1 and z_2 of g with no fixed point between them. By (b) there are neighborhoods U_1 and U_2 of z_1 and z_2, respectively, such that the orbit of each point in U_i converges to z_i. Something must happen between z_1 and z_2 to separate the points whose orbits converge to z_1 from those whose orbits converge to z_2. At the very least, we can do the following:
   (c) Prove that between every pair of fixed points of g there is a point at which g is not defined.

8.3   Chaos

In 1961, Edward Lorenz, a meteorologist at the Massachusetts Institute of Technology, was attempting to simulate weather patterns on a computer using a model derived from twelve relatively complicated equations. Rather than repeat the entirety of a computer run from the previous day, he decided to begin the calculation anew using data output from partway through the previous run. However, instead of inputting the data to the six-decimal-place accuracy that

was retained by the computer, he utilized the three-decimal-place accuracy that the computer printout provided. Much to his surprise, the results of the run were entirely different from what they had been before. He was seeing sensitivity to initial conditions. Even though the equations were fixed, just a small change in the input conditions caused a major change in the outcome. This is often called the "butterfly effect," as a butterfly fluttering its wings in Hong Kong might change conditions enough to eventually cause a tornado in Texas. In other words, small changes can have dramatic impact. This is why long-term weather prediction is so difficult. In "Deterministic nonperiodic flow," his now-classic paper about this discovery, Lorenz wrote,

> When our results . . . are applied to the atmosphere . . . they indicate that prediction of the sufficiently distant future is impossible by any method, unless the present conditions are known exactly. In view of the inevitable inaccuracy and incompleteness of weather observations, precise very-long-range forecasting would seem to be non-existent.

With the aid of a computer, Lorenz was observing a phenomenon that Henri Poincaré had intuitively described sixty years earlier, as a result of his study of the three-body problem. In a 1903 essay, "Science and Method," Poincaré suggested,

> If we knew exactly the laws of nature and the situation of the universe at the initial moment, we could predict exactly the situation of that same universe at a succeeding moment. But . . . we could still only know the initial situation *approximately*. If that enabled us to predict the succeeding situation with *the same approximation*, that is all we require. . . . But it is not always so; it may happen that small differences in the initial conditions produce very great ones in the final phenomena. [Pet]

Between the times of Poincaré and Lorenz, others had also glimpsed this intriguing phenomenon, but it took the introduction of the computer as an experimental tool in mathematics and science for it to be seen and understood broadly enough to be recognized as a robust part of many physical and mathematical systems.

A mathematical and scientific revolution was born, and over the subsequent decades many scientists and mathematicians worked to identify, describe, and define this phenomenon, its properties, and its consequences. Chaos is the name that is widely used for the general structure and behaviors that result in sensitive dependence on initial conditions.

Along with the unpredictability of sensitive dependence on initial conditions, chaos also includes an element of regularity. Orbits in a chaotic system can appear periodic over a span of time (possibly very long), but eventually they diverge into another realm of behavior, perhaps a different apparent periodicity. The regularity of the Earth orbiting the sun might be an approximate long-term periodicity in an overall chaotic system. (See [Pet] for a further discussion about this possibility.)

In this section we present a topological definition of chaos, and we discuss some essential features of chaos: it is deterministic, but unpredictable; it possesses hidden regularity, and at the same time mixes together all regions of the domain. We also examine the tent function further and present two approaches to showing that it is chaotic. Finally, we introduce sensitive dependence on initial conditions and present a theorem showing that every continuous chaotic function on an infinite metric space has sensitive dependence on its initial conditions.

DEFINITION 8.10. *Let X be a topological space. A function $f : X \to X$ is said to be **chaotic** or to **have chaos** if*

> *(i) The set of periodic points of f is dense in X,*

> *(ii) For every U, V open in X, there exists $x \in U$ and $n \in \mathbb{Z}_+$ such that $f^n(x) \in V$.*

The first condition indicates that there is regular periodic behavior densely distributed throughout the domain. No matter what point we choose in the domain, there are periodic points arbitrarily close by.

The second condition, referred to as topological transitivity, indicates that every pair of regions in the domain is mixed together by the system. Given any pair of open sets, there is at least one point in the first set that, on some iteration, is mapped into the second set.

Consider again the tent function

$$T(x) = \begin{cases} 2x & \text{if } x \in [0, \tfrac{1}{2}], \\ 2 - 2x & \text{if } x \in [\tfrac{1}{2}, 1]. \end{cases}$$

In Section 8.1 we presented the graphs of T and T^2. In Figure 8.14 we show the graphs of T^3 and T^4. The pattern is apparent. The graph of T^n results in a "tent" over $[\frac{j-1}{2^{n-1}}, \frac{j}{2^{n-1}}]$ for each $j = 1, 2, \ldots, 2^{n-1}$. It follows that each such interval contains two intersections of the graph of T^n with the line $y = x$. These intersection points are periodic points of T. Thus, as n gets larger and larger, the intervals $[\frac{j-1}{2^{n-1}}, \frac{j}{2^{n-1}}]$ partition $[0, 1]$ into smaller and smaller intervals, each of which contains periodic points.
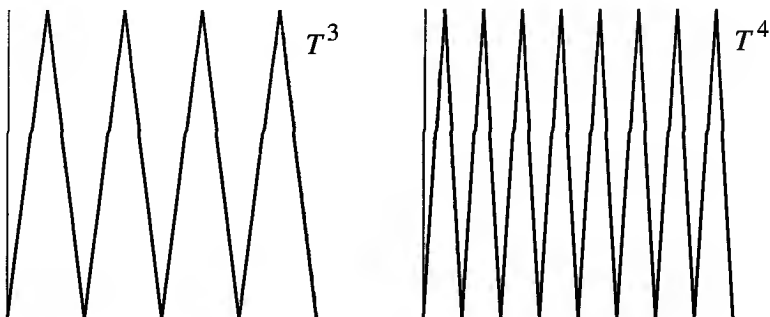
FIGURE 8 14: The graphs of T^3 and T^4.

Furthermore, T^n maps each interval $[\frac{j-1}{2^{n-1}}, \frac{j}{2^{n-1}}]$ onto $[0, 1]$. So as n gets larger and larger, smaller and smaller intervals are getting spread out or "mixed" by T^n over the whole interval $[0, 1]$. Thus the two main components of chaos appear to be present in the tent function. In a set of supplementary exercises at the end of this section, we work through the details in this approach to proving that T is chaotic.

However, we will now take a different approach to proving that T is chaotic; this approach is based on binary expansions of the real numbers in $[0, 1]$. Specifically, we use the fact that every $x \in [0, 1]$ can be expressed in the form

$$\frac{a_1}{2} + \frac{a_2}{2^2} + \ldots + \frac{a_m}{2^m} + \ldots ,$$

where each a_i equals either 0 or 1. For such x, we have the binary expansion $x = .a_1 a_2 \ldots a_m \ldots$.

For $a = 0$ or 1, we let $a^* = 1 - a$. Note that $(a^*)^* = a$. Also, if x has binary expansion $.a_1 a_2 \ldots a_m \ldots$ then $1 - x$ has binary expansion $.a_1^* a_2^* \ldots a_m^* \ldots$.

Now, if $x \in [0, \frac{1}{2}]$ then x has a binary expansion in the form $.0a_2 \ldots a_m \ldots$, and it follows that

$$2x = 2\left(\frac{0}{2} + \frac{a_2}{2^2} + \ldots + \frac{a_m}{2^m} + \ldots \right) = .a_2 a_3 \ldots a_m \ldots .$$

Furthermore, if $x \in [\frac{1}{2}, 1]$, then x has a binary expansion in the form $.1a_2 \ldots a_m \ldots$; so $1 - x = .0a_2^* \ldots a_m^* \ldots$, and $2 - 2x = .a_2^* a_3^* \ldots a_m^* \ldots$. (Note that binary expansions are not unique. For example, 1/2 can be expressed as both $.1000000 \ldots$ and $.01111111 \ldots$, and for that reason we include 1/2 in each of the possibilities, $x \in [0, \frac{1}{2}]$ or $x \in [\frac{1}{2}, 1]$, presented here.)

Therefore, using binary expansions, we can express the tent function $T : [0, 1] \to [0, 1]$ as follows:

$$T(.a_1 a_2 \ldots a_m \ldots) = \begin{cases} .a_2 a_3 \ldots a_m \ldots & \text{if } a_1 = 0, \\ .a_2^* a_3^* \ldots a_m^* \ldots & \text{if } a_1 = 1. \end{cases}$$

For example, $T(.000 \ldots) = .000 \ldots$, confirming our fixed point at 0. The binary expansion for 1 is $.111 \ldots$, and $T(.111 \ldots) = .000 \ldots$, so $T(1) = 0$. For 2/3, the binary expansion is $.1010 \ldots$, and $T(.1010 \ldots) = .1010 \ldots$; thus (as we knew already) 2/3 is a fixed point. For 2/5 we have binary expansion $.01100110 \ldots$, and $T(.01100110 \ldots) = .11001100 \ldots = 4/5$. On the other hand, $T(.11001100 \ldots) = .011011 \ldots = 2/5$. Therefore, as we saw earlier, we have period-2 points at 2/5 and 4/5.

With this formulation of T, we also have a simply way to express the value of $T^n(.a_1 a_2 \ldots a_m \ldots)$, as the following lemma indicates:

LEMMA 8.11. *For* $.a_1 a_2 \ldots a_m \ldots \in [0, 1]$ *and* $n \in \mathbb{Z}_+$,

$$T^n(.a_1 a_2 \ldots a_m \ldots) = \begin{cases} .a_{n+1} a_{n+2} \ldots & \text{if } a_n = 0, \\ .a_{n+1}^* a_{n+2}^* \ldots & \text{if } a_n = 1. \end{cases}$$

Proof. We prove this by induction on n. For $n = 1$, the result follows by the definition of T. Assume that the result holds for $n - 1$. Then

$$T^n(.a_1 a_2 \ldots a_m \ldots) = T(T^{n-1}(.a_1 a_2 \ldots a_m \ldots))$$
$$= \begin{cases} T(.a_n a_{n+1} \ldots) & \text{if } a_{n-1} = 0, \\ T(.a_n^* a_{n+1}^* \ldots) & \text{if } a_{n-1} = 1, \end{cases}$$

where the second equality holds by the inductive hypothesis. Assume that $a_n = 0$; then $T^n(.a_1 a_2 \ldots a_m \ldots)$ equals either $T(.0 a_{n+1} a_{n+2} \ldots)$ or $T(.1 a_{n+1}^* a_{n+2}^* \ldots)$. In either case, by the definition of T, the result is $.a_{n+1} a_{n+2} \ldots$, as desired.

Now assume that $a_n = 1$. Then $T^n(.a_1 a_2 \ldots a_m \ldots)$ equals either $T(.1 a_{n+1} a_{n+2} \ldots)$ or $T(.0 a_{n+1}^* a_{n+2}^* \ldots)$, and in either case we obtain the desired result, $.a_{n+1}^* a_{n+2}^* \ldots$.

Thus, if the result holds for $n - 1$ it holds for n, and it follows by induction that the result holds for all $n \in \mathbb{Z}_+$.   ∎

In our proof that T is chaotic, we use the following lemma, which indicates that if two points in $[0, 1]$ have associated binary expansions that agree in their first n entries, then the distance between those two points is at most $\frac{1}{2^n}$:

LEMMA 8.12. *Let* x *and* y *have binary expansions* $.a_1 a_2 \ldots a_m \ldots$ *and* $.b_1 b_2 \ldots b_m \ldots$, *respectively. If* $a_i = b_i$ *for* $i = 1, \ldots, n$, *then* $|x - y| \leq \frac{1}{2^n}$.

Proof. We have

$$|x - y| = \left| \sum_{j=n+1}^{\infty} \frac{a_j - b_j}{2^j} \right|$$
$$\leq \sum_{j=n+1}^{\infty} \left| \frac{a_j - b_j}{2^j} \right|$$
$$\leq \sum_{j=n+1}^{\infty} \frac{1}{2^j} \qquad (\text{since } a_j - b_j \leq 1)$$
$$= \frac{1}{2^n} \sum_{j=1}^{\infty} \frac{1}{2^j}$$
$$= \frac{1}{2^n} \qquad (\text{since } \sum_{j=1}^{\infty} \frac{1}{2^j} = 1).$$

■

Now we establish that T is chaotic.

THEOREM 8.13. *The tent function T is chaotic.*

> ***Proof.*** We begin by showing that periodic points of T are dense in $[0, 1]$. It suffices to show that if $x \in [0, 1]$ and $\varepsilon > 0$, then there is a periodic point p such that $|x - p| < \varepsilon$. Thus let $x \in [0, 1]$ and $\varepsilon > 0$ be arbitrary. Assume $.a_1 a_2 \ldots a_m \ldots$ is a binary expansion of x and $n \in \mathbb{Z}_+$ is large enough that $\frac{1}{2^n} < \varepsilon$. If we let
>
> $$p = .a_1 a_2 \ldots a_n 0 a_1 a_2 \ldots a_n 0 a_1 a_2 \ldots a_n 0 \ldots ,$$
>
> then by using Lemma 8.11 it follows that p is a periodic point, and Lemma 8.12 implies that $|x - p| < \varepsilon$. Thus periodic points of T are dense in $[0, 1]$.
>
> To show topological transitivity, let U and V be open in $[0, 1]$, and let $x = .a_1 a_2 \ldots a_m \ldots \in U$ be arbitrary. Since U is open, there exists $\varepsilon > 0$ such that $(x - \varepsilon, x + \varepsilon) \cap [0, 1] \in U$. Let $n \in \mathbb{Z}_+$ be large enough that $\frac{1}{2^n} < \varepsilon$. Pick $y = .b_1 b_2 \ldots b_m \ldots \in V$, and consider the point
>
> $$p = .a_1 a_2 \ldots a_n 0 b_1 b_2 b_3 \ldots .$$
>
> By Lemma 8.12, $|x - p| \leq \frac{1}{2^n} < \varepsilon$, and therefore $p \in U$. Lemma 8.11 implies that $T^{n+1}(p) = y \in V$. It follows that T is topologically transitive and therefore is chaotic. ∎

In the proof of the topological transitivity of T, the binary expansion $.b_1 b_2 \ldots b_m \ldots$ was entirely arbitrary and could be associated with any point in $[0, 1]$. What we actually proved was that, given U open in $[0, 1]$, there exists $n \in \mathbb{Z}_+$ such that T^n maps U onto $[0, 1]$. This is not a surprise, since we noted this phenomenon earlier. Specifically, for an open set U in $[0, 1]$ there is $m \in \mathbb{Z}_+$ large enough that some interval $[\frac{j-1}{2^{m-1}}, \frac{j}{2^{m-1}}]$ is contained in U. We previously discussed that T^m maps each such interval $[\frac{j-1}{2^{m-1}}, \frac{j}{2^{m-1}}]$ onto $[0, 1]$.

Although it does not play a role in the definition of chaos, we can also see that eventual fixed points of T are dense in $[0, 1]$. It follows from Lemma 8.11 that every point in $[0, 1]$ having an associated binary expansion that ends in all 0s, all 1s, or alternating 0s and 1s is an eventual fixed point. (In fact, this accounts for all of them—see Exercise 8.23.) Given x with binary expansion $.a_1 a_2 \ldots a_m \ldots$, the point with binary expansion $.a_1 a_2 \ldots a_n 000 \ldots$ is within $\frac{1}{2^n}$ of x, and therefore arbitrarily close to x there are points that are eventual fixed points.

It is straightforward to see that the periodic points of the tent function are all unstable. In fact, the following theorem indicates that all of the periodic points are unstable for every chaotic function defined on an infinite Hausdorff space. Thus, even though periodic points are densely distributed throughout the domain, we do not expect to see any of them in the long-term in such a chaotic system.

THEOREM 8.14. *Let X be an infinite Hausdorff space. If $f : X \to X$ is chaotic then every periodic point of f is unstable.*

> ***Proof.*** See Exercise 8.25. ∎

In general, a direct proof that a function is chaotic is not nearly as easy to formulate as it is for the tent function. However, topological conjugacy can help identify chaotic functions. The following theorem shows that, as we might expect, chaos is preserved under topological conjugacy. Thus, given a chaotic function, every function that is topologically conjugate to it is also chaotic.

THEOREM 8.15. *If f and g are topologically conjugate functions and f has chaos, then g has chaos.*

Proof. See Exercise 8.26.   ∎

The function $Q : [0, 1] \to [0, 1]$ defined by $Q(x) = 4x(1 - x)$ is topologically conjugate to the tent function. A topological conjugacy is given by the homeomorphism $h : [0, 1] \to [0, 1]$ defined by $h(x) = \sin^2(\frac{\pi}{2}x)$. It is straightforward to show that $h \circ T = Q \circ h$. (See Exercise 8.27.)

Thus Theorems 8.13 and 8.15 imply that Q is chaotic. Note that the graph of Q has features similar to those in the graph of T. (See Figure 8.15.) Both functions can be viewed as stretching the interval $[0, 1]$ to twice its length, and then folding the stretched interval over so that (except at $x = 1/2$) it maps two-to-one onto $[0, 1]$.
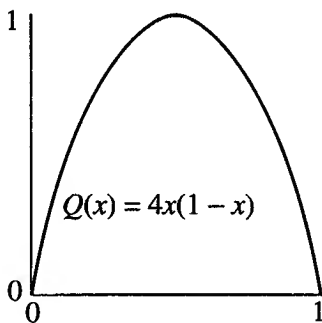
FIGURE 8.15: The function $Q(x) = 4x(1 - x)$.

The stretching and folding present in the functions T and Q is typical of the nonlinear behavior that causes chaos. The stretching tends to drive orbits apart while the folding keeps the dynamics confined to a compact domain. This combination of behaviors can result in sensitive dependence on initial conditions and the unpredictability present with chaos. Sensitive dependence on initial conditions is a metric property, defined as follows:

DEFINITION 8.16. *Let (X, d) be a metric space. A function $f : X \to X$ has **sensitive dependence on initial conditions** if there exists a $\delta > 0$ such that for every $x \in X$ and $\varepsilon > 0$, there exists $y \in B_d(x, \varepsilon)$ and $n \in \mathbb{Z}_+$ such that $d(f^n(x), f^n(y)) > \delta$.*

Thus, a function f has sensitive dependence on initial conditions if there is a distance δ such that arbitrarily close to every point x there are points whose image under some iteration of f is more that δ away from the image of x under the same iteration of f. We refer to δ as a **sensitivity constant** for f.

In Section 8.5 we will prove that chaos implies sensitive dependence on initial conditions. In the introduction to this section, we indicated that chaotic behavior is deterministic, but unpredictable, and that it possesses hidden regularity, while at the same time mixing together all regions of the domain. The hidden regularity and mixing are captured in the definition of chaos via the dense set of periodic points and topological transitivity, respectively. Being deterministic is a feature of all dynamical systems; the idea is that if you know the starting point, then the future is completely determined by the function defining the dynamical system. The unpredictability in a chaotic system comes as a result of the sensitive dependence on initial conditions. If there is any uncertainty or error in the knowledge of the initial state, then—as Lorenz suggested about the atmosphere—precise long-term predictions are not possible.

Exercises for Section 8.3

8.23. Prove that $x \in [0, 1]$ is an eventual fixed point of the tent function if and only if the binary expansion of x ends in a sequence of all 0s, all 1s, or alternating 0s and 1s.

8.24. (a) Show that if X has the trivial topology, then $f : X \to X$ is chaotic if and only if f has a periodic point.
(b) Show that if X has the discrete topology and $f : X \to X$ is chaotic, then X is finite and f is a cyclic permutation of the elements of X.

8.25. Prove Theorem 8.14: Let X be an infinite Hausdorff space. If $f : X \to X$ is chaotic then every periodic point of f is unstable.

8.26. Prove Theorem 8.15: If f and g are topologically conjugate functions and f has chaos, then g has chaos.

8.27. Verify that the homeomorphism $h : [0, 1] \to [0, 1]$, defined by $h(x) = \sin^2(\frac{\pi}{2}x)$, is a topological conjugacy between the tent function and the function $Q : [0, 1] \to [0, 1]$, defined by $Q(x) = 4x(1 - x)$.

8.28. Show that the functions $Q : [0, 1] \to [0, 1]$, defined by $Q(x) = 4x(1 - x)$, and $R : [-1, 1] \to [-1, 1]$, defined by $R(x) = 1 - 2x^2$, are topologically conjugate. (It follows R is chaotic.)

8.29. Show that the function $g : [0, 1] \to [0, 1]$, defined by

$$g(x) = \begin{cases} 3x & \text{for } 0 \le x \le \frac{1}{3}, \\ 2 - 3x & \text{for } \frac{1}{3} \le x \le \frac{2}{3}, \\ 3x - 2 & \text{for } \frac{2}{3} \le x \le 1, \end{cases}$$

is chaotic, by using ternary expansions to show each of the following:
(a) Periodic points are dense.
(b) g is topologically transitive.

8.30. Show that if $f : X \to X$ has a point x whose orbit is dense in X then f is topologically transitive. (The converse of this theorem is also true if X is a compact subset of either $\mathbb{R}$ or S^1, but the proof requires tools not introduced in this text.)

8.31. On $S^1 \subset \mathbb{R}^2$, let θ represent the point at angle θ measured counterclockwise from the positive x-axis. Define $f : S^1 \to S^1$ by $f(\theta) = 2\theta$. In (a) and (b) we prove that f is chaotic.

(a) Prove that for $n, j \in \mathbb{Z}_+$, the points with angular representation $\frac{2\pi j}{2^n - 1}$ are fixed points of f^n, and thereby show that periodic points of f are dense in S^1.

(b) Prove that every interval of the form $[\theta, \theta + \frac{\pi}{2^{n-1}}] \subset S^1$ is mapped onto S^1 by f^n, and then use this result to show that f is topologically transitive.

Supplementary Exercises: Another Approach to Proving that the Tent Function is Chaotic

We now provide a second proof that the tent function T is chaotic. This approach is easy to visualize given the nature of the graph of T^n. We begin with a lemma asserting that the graph of T^n is a "tent" over each interval $[\frac{j-1}{2^{n-1}}, \frac{j}{2^{n-1}}]$.

LEMMA 8.17. *For each $n \in \mathbb{Z}_+$ and $j = 1, 2, \ldots, 2^{n-1}$, we have*

$$
T^n(x) = \begin{cases} 2^n x - 2j + 2 & \text{if } x \in \left[\dfrac{j-1}{2^{n-1}}, \dfrac{j - 1/2}{2^{n-1}} \right], \\ -2^n x + 2^n + 2j - 2 & \text{if } x \in \left[\dfrac{j - 1/2}{2^{n-1}}, \dfrac{j}{2^{n-1}} \right]. \end{cases}
$$

In particular, over $[\frac{j-1}{2^{n-1}}, \frac{j}{2^{n-1}}]$ the function T^n equals 0 at the endpoints, equals 1 at the midpoint, and is linear on each half of the interval.

SE 8.32. Prove Lemma 8.17. (Hint: Use induction; the $n = 1$ case holds by the definition of T.)

With the lemma, it is straightforward to prove the desired theorem:

THEOREM 8.18. *The tent function T is chaotic*

> **Proof.** We begin by showing that periodic points are dense in $[0, 1]$. It suffices to prove that every open set in $[0, 1]$ contains a periodic point. Let $U \subset [0, 1]$ be an arbitrary open set.   ∎

SE 8.33. Prove that there exist $n, j \in \mathbb{Z}_+$ such that the interval $[\frac{j-1}{2^{n-1}}, \frac{j}{2^{n-1}}]$ lies in U.

SE 8.34. With n as in the previous exercise, prove that T^n has a fixed point in U and therefore that there is a periodic point of T in U.

Next we establish topological transitivity. Let U and V be arbitrary open sets in $[0, 1]$.

SE 8.35. Use Exercise SE 8.33 and Lemma 8.17 to show that there exists $n \in \mathbb{Z}_+$ such that $V \subset T^n(U)$, and thereby conclude that T is topologically transitive.

8.4   *A Simple Population Model with Complicated Dynamics*

The title of this section is an adaptation of the title of a paper, "Simple mathematical models with very complicated dynamics," by physicist and mathematical biologist Robert May, published in the journal *Nature* in 1976. May's paper was a landmark in the chaos revolution. Along with a presentation of interesting mathematical results and their potential applied consequences, this paper was a call to the scientific community to realize that simple nonlinear equations can possess a wide spectrum of behaviors from fixed point to periodic to chaotic, and that unpredictable behavior in a modeled system, previously thought to be "noise" arising externally from the environment, could be a result of chaotic behavior internal to the system. May further urged that these ideas be introduced early in students' mathematics-education experiences, "so that students' intuition may be enriched by seeing the wild things that simple nonlinear equations can do." In this section, we examine the logistic population model discussed by May and highlight some of the features of the model's behavior.

Assume that we have a species whose population from generation to generation is given by a function F, where $F(p)$ represents the population in the next generation resulting from a population of p in the current generation. We assume that the population in the next generation is directly proportional to two factors: the population in the current generation and the amount of room in the environment in which the population can grow. Thus $F(p) = kpR(p)$ where $R(p)$ is a function reflecting the room in which the population can grow. We assume that the environment can hold a maximum population of M and that $R(p)$ is simply given by $M - p$, the difference between the maximum population and the current population. Thus $F(p) = kp(M - p)$.

To simplify matters, instead of working with the variable p, we work with $x = p/M$, the fraction of the maximum possible population represented by p. Consequently, our model takes on the form $f(x) = \alpha x(1 - x)$, with domain $[0, 1]$. We call this the **logistic growth function**. We think of the parameter α as a growth rate, and we distinguish different functions with different growth rates by writing them as f_α, rather than f.

We are only interested in growth rates α such that f_α maps $[0, 1]$ back into $[0, 1]$. This occurs for $\alpha \in [0, 4]$. By the **logistic family**, we mean for the family of logistic functions corresponding to parameter values $[0, 4]$. We sketch the graphs of f_0, f_1, f_2, f_3, and f_4 in Figure 8.16. For f_0 the dynamics are relatively simple: no matter what the population is in one generation, it
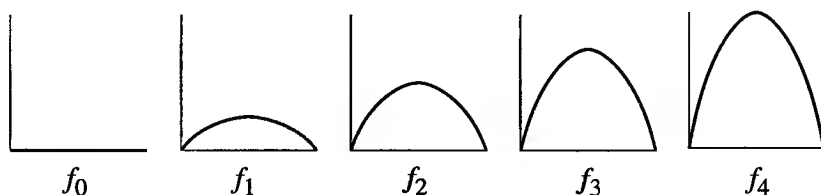
FIGURE 8.16: The logistic functions f_0, f_1, f_2, f_3, and f_4.

goes immediately to 0 in the next. For values of α near 0, the dynamics of f_α are not much different. Specifically, such f_α have an asymptotically stable fixed point at 0, and every orbit in [0, 1] approaches 0 in the limit. Thus, for small values of α the growth rate is not large enough to sustain a nonzero population, and initial nonzero populations go to 0 in the limit. On the other hand, at the other end of our parameter domain where $\alpha = 4$, we have the function $f_4(x) = 4x(1 - x)$, which we showed to be chaotic in the previous section.

Thus we are left with the question, how do the dynamics of f_α change as we go from a function f_0, where all points immediately map to 0, to a function f_4 that is chaotic? The road to chaos is a fascinating one. Unfortunately, it cannot be analyzed, or even described, in just a few pages. We discuss only some of the highlights seen along the way. Further details can be found in many of the introductory texts on dynamical systems and chaos and in other references we will point out.

To begin, consider Figure 8.17. We call it an **orbit diagram**; it depicts how the expected long-term behavior of the functions f_α evolves as α transitions from 0 to 4. In the diagram, the horizontal axis represents the parameter α, and the vertical axis represents the variable x in the domain, [0, 1], of each of the functions f_α. The diagram is generated by setting α equal to each of the 9601 values, $0, \frac{1}{2400}, \frac{2}{2400}, \ldots, \frac{9599}{2400}, 4$, and doing the following for each:

(i) Compute the first 2075 points in the orbit of 0.5 under f_α; that is, compute

$$f_\alpha(0.5), f_\alpha^2(0.5), \ldots, f_\alpha^{2075}(0.5).$$

(ii) In the diagram, plot the following points corresponding to the last 75 of the 2075 computed orbit points:

$$(\alpha, f_\alpha^{2001}(0.5)), \ldots, (\alpha, f_\alpha^{2075}(0.5)).$$
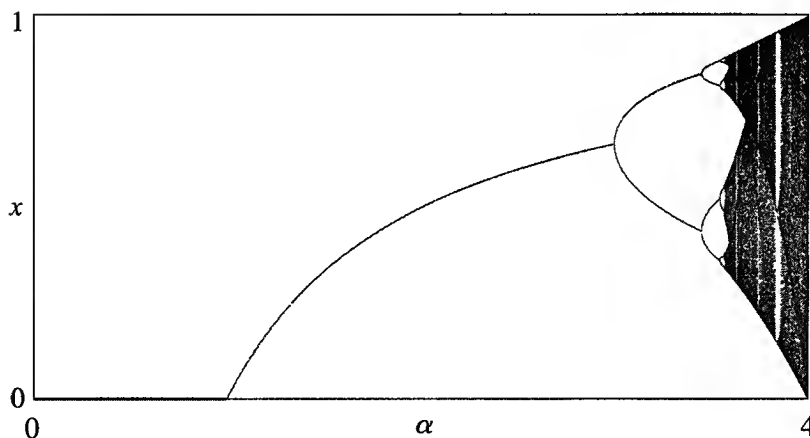
FIGURE 8.17: The orbit diagram for the logistic family. (Courtesy of Chip Ross.)

Thus for each $\alpha \in [0, 4]$, in the vertical slice at α in the diagram, we are seeing an estimate of the long-term behavior of f_α. For example, at $\alpha = 1$, just the value 0 is plotted. Therefore for f_1, the expected long-term behavior is to approach a fixed point at 0. For $\alpha = 2$, a single point at 0.5 is plotted. Thus for f_2 the expected long-term behavior is to approach a fixed point at 0.5. Also, for $\alpha = 3.4$, values at approximately 0.45 and 0.85 are plotted. This reflects that for $f_{3.4}$, the expected long-term behavior is to approach a periodic orbit oscillating between approximately 0.45 and 0.85. Let us verify these observations analytically and further explore the behavior depicted in the orbit diagram.

To determine the fixed points for f_α, we solve $f_\alpha(x) = x$. We find solutions at 0 and at $\frac{\alpha-1}{\alpha}$. Note that $\frac{\alpha-1}{\alpha}$ is outside our domain $[0, 1]$ for $\alpha \in [0, 1]$, but inside the domain for $\alpha \in [1, 4]$. Using Theorem 8.9 to analyze the stability of the fixed points, we see that 0 is asymptotically stable for $\alpha \in [0, 1)$ and unstable for $\alpha \in (1, 4]$, and $\frac{\alpha-1}{\alpha}$ is asymptotically stable for $\alpha \in (1, 3)$ and unstable for $\alpha \in (3, 4]$. (See Exercise 8.36.) In Figure 8.18, we plot these fixed points on an x versus α coordinate system similar to that used in the orbit diagram. We use a solid curve to indicate the values where the fixed point is asymptotically stable and use a dashed curve to indicate the values where the fixed point is unstable.
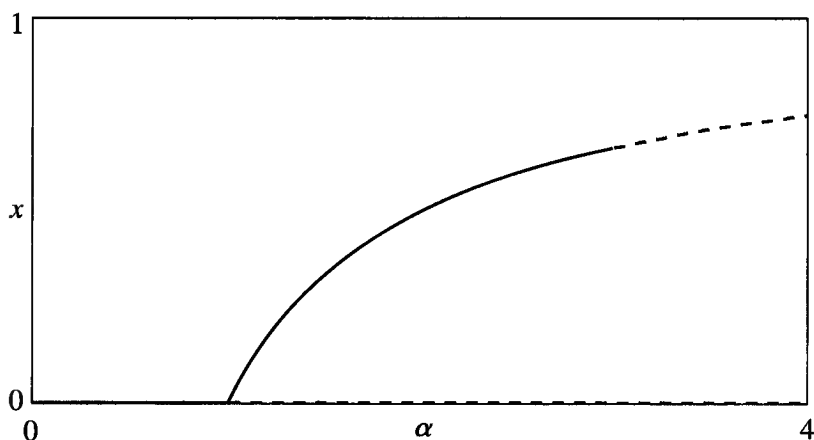
FIGURE 8.18: The fixed points of the logistic family and their stability.

The value of α where the fixed point 0 loses its stability is exactly the value of α where the fixed point $\frac{\alpha-1}{\alpha}$ merges with the fixed point at 0 and enters our domain. Furthermore, the fixed point $\frac{\alpha-1}{\alpha}$ appears to lose its stability at approximately the location where the period-2 orbit comes into our picture in the orbit diagram. We explore the birth of the period-2 orbit next. The changes in stability we see here are examples of behavior changes known as **bifurcations**.

To find period-2 points for f_α, we solve $f_\alpha^2(x) = x$. Along with the fixed points of f_α, we also find solutions at

$$\frac{1 + \alpha \pm \sqrt{\alpha^2 - 2\alpha - 3}}{2\alpha}.$$

These are period-2 points, and for each α they form a period-2 orbit. The period-2 points are defined for $\alpha \in (3, 4]$, and they emerge in our domain exactly where we thought they would, at the value of α where the fixed point $\frac{\alpha-1}{\alpha}$ loses its stability. At $\alpha = 3$ the fixed point $\frac{\alpha-1}{\alpha}$ has undergone what is known as a **period-doubling bifurcation**; it has given up its stability to an emerging period-2 orbit. If we use Theorem 8.9 to determine the stability of these period-2 points, we find that they are asymptotically stable for $\alpha \in (3, 1 + \sqrt{6})$ and unstable for $\alpha \in (1 + \sqrt{6}, 4]$. (See Exercise 8.38.) In Figure 8.19 we expand Figure 8.18 to include the period-2 points. Note the similarity between Figure 8.19 and the orbit diagram over the parameter values $0 \le \alpha \le 1 + \sqrt{6}$.
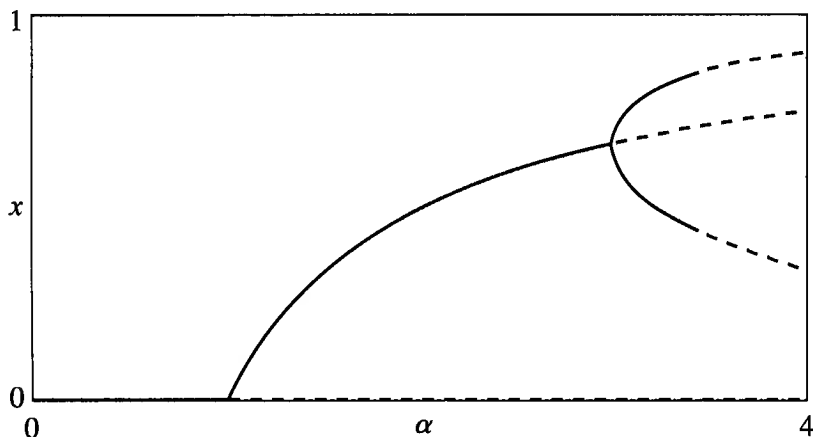
FIGURE 8.19: The fixed points and period-2 points of the logistic family.

As we observed, the period-2 points lose their stability at $1 + \sqrt{6}$. What happens there? In Figure 8.20 we zoom in on the orbit diagram from Figure 8.17, considering only the parameter values $3.4 \le \alpha \le 3.6$. Examining this diagram, it appears that an asymptotically stable period-4 orbit has been born at the point where the period-2 orbit loses its stability. Thus we have had another period doubling. It is difficult to verify this analytically by solving $f_\alpha^4(x) = x$, because that involves solving a polynomial of degree 16. However, the period doubling can be verified by examining the equations $f_\alpha^4(x) = x$ graphically for values of α near $1 + \sqrt{6}$. (See Exercise 8.40.)

We are still just at the beginning of the story behind the orbit diagram. As the parameter α increases, the period-4 orbit bifurcates to a period-8 orbit that bifurcates to a period-16 orbit, and so on. These parameter intervals between period doublings get shorter and shorter, and the parameter values associated to the period doublings approach a limit $L \approx 3.57$.
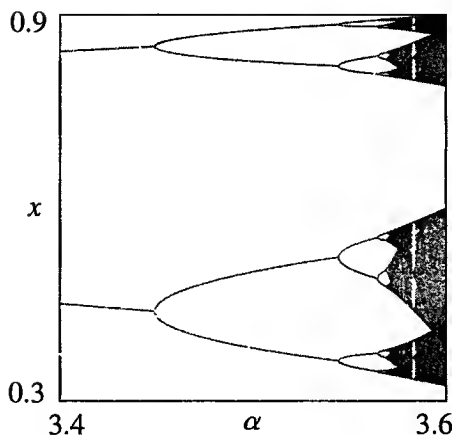
FIGURE 8.20: The orbit diagram for f_α with $3.4 \leq \alpha \leq 3.6$. (Courtesy of Chip Ross.)

At the period-doubling limit L we have infinitely many periodic orbits, and we enter a region of the parameter domain where a variety of interesting behaviors are seen. Again, we cannot provide all of the details, but here are some of the highlights:

(i) There are infinitely many parameter values α where f_α is chaotic on an infinite subset of $[0, 1]$.

(ii) For each $\alpha \in [L, 4]$, f_α has infinitely many periodic points.

(iii) There is an open dense subset P of $[L, 4]$ such that for all $\alpha \in P$, f_α has exactly one asymptotically stable periodic orbit. For all other $\alpha \in [L, 4]$, there are no asymptotically stable periodic orbits of f_α.

(iv) For each odd n, there is an open interval of parameter values where there is an asymptotically stable period-n orbit.

(v) Each of the just mentioned period-n orbits arises out of what is known as a tangent bifurcation (see Exercise 8.42) and eventually loses its stability at the beginning of a cascade of period doublings similar to that previously observed.

In the orbit diagram, we can see "windows" where the asymptotically stable orbits with odd period exist and where their subsequent period doublings take place. For example, in Figure 8.21 we examine more closely the part of the orbit diagram corresponding to $3.8 \leq \alpha \leq 3.9$. We can see that the period-3 orbit enters at approximately $\alpha = 3.83$ and then bifurcates to a period-6 orbit at approximately $\alpha = 3.84$. The birth of the period-3 orbit is addressed further in Exercise 8.43.
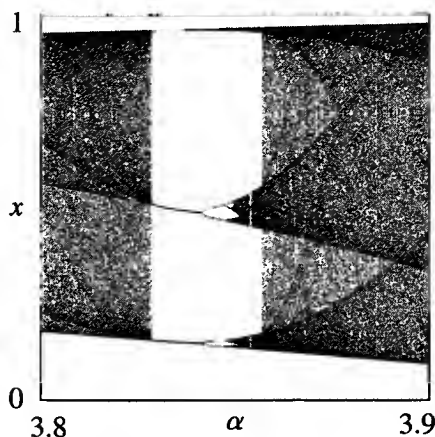
FIGURE 8.21: The orbit diagram for f_α with $3.8 \le \alpha \le 3.9$, showing the birth of a period-3 orbit. (Courtesy of Chip Ross.)

The logistic family and others like it have been extensively studied to uncover the dynamic story behind the orbit diagram. (See [DevR] and [Gil], for example.) One fascinating result is that the spectrum of behaviors we have described is not unique to the logistic family, but can be seen in other families of functions defined over an interval where the family rises through a collection of one-humped functions to a function that fills out the interval with chaos. The family of functions $s_\alpha(x) = \alpha \sin(\pi x)$, defined on $[0, 1]$, with α going from 0 to 1, is one such family.

Now, what does all of this say about the population being modeled by the logistic equation? A main theme of May's paper was that even very simple functions can exhibit complicated dynamics, and therefore the complex behaviors observed in simple natural systems could be inherent to the systems, rather than a result of influences external to them. Fixed points, periodic points, and chaos are all phenomena to be expected in natural systems, and as parameters change in the systems, we can expect exotic transitions between these phenomena similar to those observed in the logistic family.

As the chaos revolution took hold, many mathematicians and scientists began to examine systems for the behaviors seen in the logistic family. The dynamics in this simple model have helped deepen the understanding of phenomena such as turbulence in fluid flow, motions of celestial bodies, outbreaks of diseases, and vibrations in machinery.

Chaos is not the only source of complicated, unpredictable behavior, but the identification of chaos and its associated structures has helped mathematicians and scientists better understand the dynamics of nonlinear systems. And, of course, topological concepts have played an important role along the way.

Exercises for Section 8.4

In these exercises, f_α refers to the functions defined by $f_\alpha(x) = \alpha x(1 - x)$ with $\alpha \in [0, 4]$.

8.36. Determine the fixed points of f_α, and use Theorem 8.9 to deduce their stability.

8.37. Over the domain $x \in [-1, 1]$ sketch separate web diagrams for f_α with α slightly less than 1, equal to 1, and slightly greater than 1. Discuss the changes you observe in the fixed points and their stability as the parameter α passes through 1. The bifurcation observed here is known as a **transcritical bifurcation**.

8.38. Determine the period-2 points of f_α, and use Theorem 8.9 to deduce their stability. (Hint: Finding the period-2 points involves solving a degree-4 polynomial; the fixed points of f_α are two of the solutions to the polynomial.)

8.39. Using a computer graphing program, examine the graphs of the function f_α^2 and the line $y = x$ for α slightly less than 3, equal to 3, and slightly greater than 3 in order to approximate the fixed points of f_α^2. Which fixed points of f_α^2 that you observe are fixed points of f_α and which are period-2 points of f_α? Discuss the changes you observe as the parameter α passes through 3. The bifurcation observed here is known as a **period-doubling bifurcation**.

8.40. Using a computer graphing program, examine the graphs of f_α^4 and the line $y = x$ in order to approximate the value of α where the period-doubling bifurcation, from an asymptotically stable period-2 orbit to an asymptotically stable period-4 orbit, takes place.

8.41. Consider the family of functions $g_\lambda : \mathbb{R} \to \mathbb{R}$ given by $g_\lambda(x) = \lambda x - x^3$ with $\lambda \in [0, 2]$.
  (a) Determine the fixed points and their stability.
  (b) Discuss the bifurcation that takes place over the parameter domain, and include illustrations showing how the graph of g_λ changes (in relation to the line $y = x$) as the bifurcation takes place. This bifurcation is called a **pitchfork bifurcation**. It is a pitchfork bifurcation that occurs at $\alpha = 3$ in f_α^2 that causes the first period-doubling bifurcation in the logistic family.

8.42. Consider the family of functions $h_\lambda : \mathbb{R} \to \mathbb{R}$ given by $h_\lambda(x) = \lambda + x - x^2$ with $\lambda \in [-1, 1]$.
  (a) Determine the fixed points and their stability.
  (b) Discuss the bifurcation that takes place over the parameter domain, and include illustrations showing how the graph of h_λ changes (in relation to the line $y = x$) as the bifurcation takes place. This bifurcation is called a **tangent bifurcation**. Explain the name.

8.43. Using a computer graphing program, examine the graphs of f_α^3 and the line $y = x$ in order to approximate the value of α where the tangent bifurcation takes place that gives rise to the period-3 window shown in Figure 8.21.

8.5   *Chaos Implies Sensitive Dependence on Initial Conditions*

In this section we prove that the metric notion of sensitive dependence on initial conditions is a consequence of the topological definition of chaos. Specifically, we have the following theorem:

THEOREM 8.19. *Let X be an infinite metric space and $f : X \to X$ be continuous and chaotic. Then f has sensitive dependence on initial conditions.*

This result was proven by John Banks, Jeff Brooks, Grant Cairns, Gary Davis, and Peter Stacey in 1992 in [Ban]. In some of the first formulations of chaos, sensitive dependence on initial conditions was included among the requirements for a function to be chaotic. (For example, see [DevR].) The

realization that sensitive dependence on initial conditions is a consequence of the two topological requirements for chaos made it possible to formulate a definition of chaos solely in topological terms.

> **Proof.** Let X be an infinite metric space with metric d, and assume that $f : X \to X$ is continuous and chaotic. We show that there exists a $\delta > 0$ such that for all $x \in X$ and $\varepsilon > 0$, there exist $y \in B_d(x, \varepsilon)$ and $n \in \mathbb{Z}_+$ such that $d(f^n(x), f^n(y)) > \delta$.
>
> To begin, choose two periodic points q and q' whose orbits are disjoint. Let δ_0 be the minimum distance between the orbits (that is, δ_0 is the minimum distance between any two points, one on the orbit of q, the other on the orbit of q'). Since the orbits are disjoint, $\delta_0 > 0$. Set $\delta = \delta_0/8$. We claim that δ satisfies the requirements of the theorem. Thus assume that $x \in X$ is arbitrary and $\varepsilon > 0$. Without loss of generality, we may assume that $\varepsilon \leq \delta$. We show that there exist $y \in B_d(x, \varepsilon)$ and $n \in \mathbb{Z}_+$ such that $d(f^n(x), f^n(y)) > \delta$.
>
> Since periodic points of f are dense, there exists a periodic point p in $B_d(x, \varepsilon)$. Assume that p has period m.
>
> Either the distance from x to every point in the orbit of q is greater than $\delta_0/2$ or the distance from x to every point in the orbit of q' is greater than $\delta_0/2$. Otherwise, by the triangle inequality, there is a pair of points—one in the orbit of q and the other in the orbit of q'—whose distance apart is less than δ_0, contradicting the definition of δ_0. Without loss of generality, assume that the distance from x to every point in the orbit of q is greater than $\delta_0/2 = 4\delta$.
>
> Let B_0 be the open ball about q of radius δ, and for each $j = 1, \ldots, m$, let B_j be the open ball about $f^j(q)$ of radius δ. By the continuity of each f^j, the sets $(f^j)^{-1}(B_j)$ are open sets. Furthermore, each such set contains q. Set
>
> $$V = B_0 \cap \left(\bigcap_{j=1}^{m} ((f^j)^{-1}(B_j)) \right).$$

The set V is an open set containing q. Note that the m used here is the period of p and is not necessarily the period of q. The idea behind the construction of V is to find a neighborhood of q such that for every v in the neighborhood, the points $f^1(v), \ldots, f^m(v)$, obtained from the first m iterations of f on v, are all within δ of the orbit of q. The reason for needing this will become clear shortly.

> By topological transitivity there exist $w \in B_d(x, \varepsilon)$ and $k \in \mathbb{Z}_+$ such that $f^k(w) \in V$. The integers k and $k+m$ are m apart, so there exists $h \in \mathbb{Z}_+$ such that $k \leq hm \leq k+m$. We claim that $d(f^{hm}(p), f^{hm}(w)) > 2\delta$. This implies that the distance between $f^{hm}(x)$ and either $f^{hm}(p)$ or $f^{hm}(w)$ must be greater than δ, and since both p and w are in $B_d(x, \varepsilon)$, this shows that δ satisfies the requirements of the theorem. We return to this point in a moment.

To prove the claim that $d(f^{hm}(p), f^{hm}(w)) > 2\delta$, first note that $f^{hm}(p) = p$, so we show that $d(p, f^{hm}(w)) > 2\delta$. Also note that since $f^k(w) \in V$, the next m iterations of f on w are guaranteed (as we indicated earlier) to all lie within δ of the orbit of q. In particular, $f^{hm}(w) \in B_{hm-k}$, and therefore $d(f^{hm}(w), f^{hm-k}(q)) < \delta$. Now, by the triangle inequality,

$$
\begin{aligned}
d(x, f^{hm-k}(q)) &\le d(x, p) + d(p, f^{hm}(w)) + d(f^{hm}(w), f^{hm-k}(q)) \\
&\le \delta + d(p, f^{hm}(w)) + \delta \\
&= 2\delta + d(p, f^{hm}(w)).
\end{aligned}
$$

Since the distance between x and the orbit of q is greater than 4δ, it follows that $4\delta < d(x, f^{hm-k}(q))$. Therefore $4\delta < 2\delta + d(p, f^{hm}(w))$, implying that $d(p, f^{hm}(w)) > 2\delta$. Thus, we have established the claim that $d(f^{hm}(p), f^{hm}(w)) > 2\delta$.

Now, let $n = hm$. Either

$$
d(f^n(x), f^n(w)) > \delta \text{ or } d(f^n(x), f^n(p)) > \delta,
$$

otherwise we would have

$$
d(f^n(x), f^n(w)) \le \delta \text{ and } d(f^n(x), f^n(p)) \le \delta,
$$

which, by the triangle inequality, would imply that $d(f^n(p), f^n(w)) \le 2\delta$, a contradiction. Since both w and p lie in $B_d(x, \varepsilon)$, there exists y in $B_d(x, \varepsilon)$ and $n \in \mathbb{Z}_+$ such that $d(f^n(x), f^n(y)) > \delta$, completing the proof of Theorem 8.19. ∎

Following the results of Banks and his colleagues, others explored further relationships between the existence of a dense set of periodic points, topological transitivity, and sensitive dependence on initial conditions. A number of interesting results emerged. For example, in 1994 Michel Vellekoop and Raoul Berglund showed in [Vel] that for I, an interval in $\mathbb{R}$, if $f : I \to I$ is continuous and has topological transitivity, then f has a dense set of periodic points and therefore is chaotic. Thus, for a continuous function on an interval in $\mathbb{R}$, topological transitivity suffices to assert the existence of chaos.

Also, in 1997 Pat Touhey showed in [Tou] that for continuous $f : X \to X$, the two conditions for chaos in Definition 8.10 can be consolidated into one necessary condition via the following theorem:

THEOREM 8.20. *A continuous function $f : X \to X$ is chaotic if and only if for every pair of open sets U and V in X there is a periodic point $x \in U$ and a $k \in \mathbb{Z}_+$ such that $f^k(x) \in V$.*

Proof. See Exercise 8.47. ∎

The condition for a function f to be chaotic, specified in Theorem 8.20, states that every pair of open sets in the domain of f is visited by some periodic orbit of f. Touhey further showed that given a finite collection of open sets in the domain of a chaotic function f, there is a periodic orbit of f that visits each of the open sets.

Although Theorem 8.20 allows for a simplification of the definition of chaos to one necessary condition (in the case of a continuous function), the definition with two necessary conditions (Definition 8.10) is particularly revealing because it highlights the dyadic nature of chaos: the regularity present in the dense collection of periodic points and the irregularity resulting from topological transitivity.

Exercises for Section 8.5

8.44. Here we show that neither topological transitivity nor the existence of a dense set of periodic points is by itself enough to imply sensitive dependence on initial conditions.

   (a) Find a function $f : X \to X$ that have a dense set of periodic points in its domain but does not have sensitive dependence on initial conditions.

   (b) Let $f : S^1 \to S^1$ be defined by $f(\theta) = \theta + 1$; that is, f is rotation of the circle by one radian. Prove that f is topologically transitive but does not have sensitive dependence on initial conditions. (Hint: Given an interval U in the circle, prove that there exists $n \in \mathbb{Z}_+$ such that $f^n(U) \cap U \neq \varnothing$, but $f^n(U) \neq U$. Then consider the sets $f^{mn}(U)$ for $m \in \mathbb{Z}_+$.)

8.45. In this problem, we demonstrate that sensitive dependence on initial conditions is not preserved under topological conjugacy. Let $f : (1, \infty) \to (1, \infty)$ be defined by $f(x) = 2x$, and let $g : \mathbb{R}_+ \to \mathbb{R}_+$ be defined by $g(x) = 1 + x$.

   (a) Prove that f has sensitive dependence on initial conditions, but g does not.

   (b) By finding an explicit homeomorphism $h : (1, \infty) \to \mathbb{R}_+$, prove that f and g are topologically conjugate.

8.46. Exercise 8.45 demonstrates that sensitive dependence on initial conditions is not generally preserved under topological conjugacy, but, as we show in this exercise, if we restrict to a compact domain, then it is.

   Let X and Y be compact metric spaces with metrics d_X and d_Y, respectively. Assume that $f : X \to X$ has sensitive dependence on initial conditions with sensitivity constant δ, and assume that $g : Y \to Y$ is topologically conjugate to f via a homeomorphism $h : X \to Y$.

   (a) Prove that $C = \{(y_1, y_2) \in Y \times Y \mid d_X(h^{-1}(y_1), h^{-1}(y_2)) > \delta\}$ is a compact subset of $Y \times Y$.

   (b) Prove that $d_Y(y_1, y_2) > 0$ for all $(y_1, y_2) \in C$ and therefore d_Y takes on a minimum positive value δ^* on C.

   (c) Prove that g has sensitive dependence on initial conditions with sensitivity constant δ^*.

8.47. Prove Theorem 8.20: A continuous function $f : X \to X$ is chaotic if and only if for every pair of open sets U and V in X there is a periodic point $x \in U$ and a $k \in \mathbb{Z}_+$ such that $f^k(x) \in V$.

Homotopy and Degree Theory

Although we have discussed how we might compare two different topological spaces, we have not talked about how we might compare two functions. Is there a useful equivalence for functions? In Section 9.1, we introduce the concept of homotopy, a continuous deformation of one function to another. Properties shared by two functions that are homotopic to each other play an important role in the results we present in this chapter.

In Section 9.2, we apply homotopy to functions we call circle functions. The circle is a natural space to use when modeling situations where a variable is periodic that is, repeating the same behavior over successive time intervals of a fixed length. Circle functions arise in situations where one periodic variable is being mapped to another. For example, when a beating heart undergoes a stimulus, the time in the beat cycle at which the stimulus occurs is mapped to the time in the beat cycle when the heart recovers from the stimulus and beats again.

Circle functions are continuous functions mapping the circle to itself. For such functions we introduce the notion of degree, a quantity that measures how many times the function wraps the circle around itself. It is a straightforward consequence of our definition of degree that two circle functions are homotopic if and only if they have the same degree. We define degree in Section 9.2, and in Section 9.6 we prove some technical results behind the definition of degree.

In Section 9.2, we also introduce retractions and use degree to prove the Two-Dimensional No Retraction Theorem (a result that we use in Chapter 10 to prove the Two-Dimensional Brouwer Fixed Point Theorem). In Section 9.3 we present an application of degree to a heartbeat model. There, with simple assumptions about a beating heart, we draw the conclusion that the heart cannot respond continuously to a stimulus applied at varying strengths and times in its beat cycle. We discuss the surprising conclusion that a relatively small stimulus to the heart at just the right moment in the beat cycle may throw the heart into fibrillation.

In this chapter, we also consider some applications of degree in mathematics. In Section 9.4 we use degree to prove the Fundamental Theorem of Algebra. In Section 9.5 we use degree to prove that $\mathbb{R}^2 - \{O\}$ and S^1 are not simply connected, then use those results to distinguish the plane from $\mathbb{R}^n$, for $n \geq 3$, and to distinguish the circle from S^n, for $n \geq 2$.

9.1   Homotopy

Here we introduce homotopy, a concept that makes precise what it means to deform one function to another.

DEFINITION 9.1. *Let* $f, g : X \to Y$ *be continuous functions. Assume that* $I = [0, 1]$ *has the subspace topology it inherits from* $\mathbb{R}$ *and that* $X \times I$ *has the product topology. We say that* f *and* g *are* **homotopic** *if there exists a continuous function* $F : X \times I \to Y$ *such that* $F(x, 0) = f(x)$ *and* $F(x, 1) = g(x)$. *Such a function* F *is called a* **homotopy** *between* f *and* g. *The expression* $f \simeq g$ *denotes that* f *and* g *are homotopic.*

Figure 9.1 depicts a homotopy F. With $X \times I$ having the product topology, we think of a homotopy as an interval, I, of maps from X to Y. For a fixed value t in I, the function $F|_{X \times \{t\}}$ is a map from X to Y. As the values for t vary over the interval I, the map varies continuously over a one-parameter family of continuous maps, starting with f and ending with g.
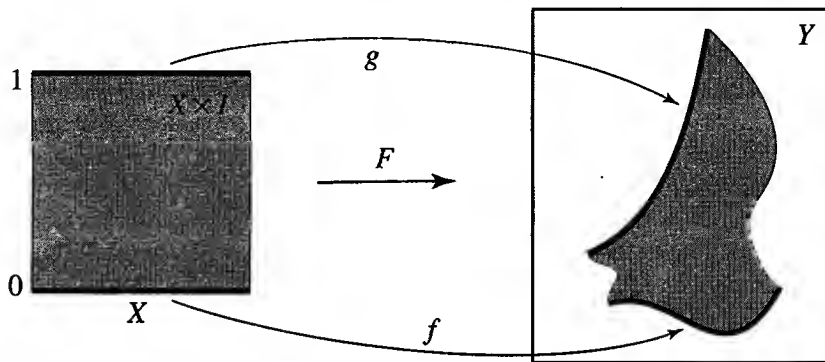
FIGURE 9.1: A homotopy deforms map f to map g.

EXAMPLE 9.1. Define $F : \mathbb{R} \times I \to \mathbb{R}$ by $F(x, t) = x + t$. The function F is continuous since addition is continuous. So F is a homotopy between $f(x) = F(x, 0) = x$ and $g(x) = F(x, 1) = x + 1$. The map f is the identity map on $\mathbb{R}$, sending each point in $\mathbb{R}$ to itself. The map g translates the entire real line one unit in the positive direction. For a fixed value of t, the homotopy $F(x, t)$ translates the real line a distance t. As t varies from 0 to 1, F takes us from f to g.

EXAMPLE 9.2. Define $f : I \to \mathbb{R}^2$ by $f(x) = (\cos(\pi x), -\sin(\pi x))$ and $g : I \to \mathbb{R}^2$ by $g(x) = (\cos(\pi x), \sin(\pi x))$. Both of these maps are examples of paths.

They are homotopic, but how do we see the homotopy? We take a straight-line homotopy between them. We call it a straight-line homotopy since for each value of $x \in I$, we deform $f(x)$ to $g(x)$ along the line segment between them. (See Figure 9.2.) During the homotopy, at each t between 0 and 1, we have moved that percentage of the way from $f(x)$ to $g(x)$.

Define $F : I \times I \to \mathbb{R}^2$ by $F(x, t) = (\cos(\pi x), (1 - 2t) \sin(\pi x))$. The function $F(x, t)$ is continuous since it is made up of compositions and products of continuous functions. Furthermore, $F(x, 0) = f(x)$ and $F(x, 1) = g(x)$, and therefore F is the desired homotopy.
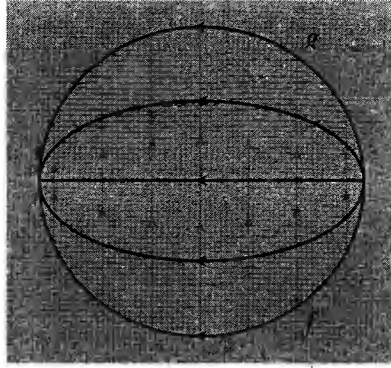
FIGURE 9.2: A homotopy in $\mathbb{R}^2$ from path f to path g.

In the previous example, if we considered the two paths f and g to have range $\mathbb{R}^2 - \{O\}$, then the straight-line homotopy between them is no longer valid, since the function F, as defined, does not send $I \times I$ to $\mathbb{R}^2 - \{O\}$. However, even though F does not work as a homotopy, the two paths are homotopic in $\mathbb{R}^2 - \{O\}$. (See Exercise 9.4.)

THEOREM 9.2. *The relation $\simeq$ is an equivalence relation on the set of all continuous functions $f : X \to Y$.*

Proof. We must show that the relation is reflexive, symmetric, and transitive. To show that $\simeq$ is reflexive, let $f : X \to Y$ be a continuous function. Define a homotopy $F : X \times I \to Y$ by $F(x, t) = f(x)$. Then $F(x, 0) = f(x)$ and $F(x, 1) = f(x)$, so $f \simeq f$.

That $\simeq$ is symmetric follows from the fact that if $F(x, t)$ is a homotopy between f and g, then $G : X \times I \to Y$, defined by $G(x, t) = F(x, 1 - t)$, is a homotopy between g and f. So $f \simeq g$ implies $g \simeq f$.

To prove transitivity, let f, g, and h be continuous maps from X to Y. Suppose that f is homotopic to g via the homotopy F and g is homotopic to h via the homotopy G. Then, as illustrated in Figure 9.3, define a function $H : X \times I \to Y$ by

$$H(x, t) = \begin{cases} F(x, 2t) & \text{for } 0 \le t \le \frac{1}{2}, \\ G(x, 2t - 1) & \text{for } \frac{1}{2} \le t \le 1. \end{cases}$$

For $t = 1/2$ the expressions $F(x, 2t)$ and $G(x, 2t - 1)$ both equal $g(x)$, and hence they agree on the set where they are both used in the definition. It follows that H is continuous by the Pasting Lemma. Since $H(x, 0) = f(x)$ and $H(x, 1) = h(x)$, we have proved that $f \simeq g$ and $g \simeq h$ imply $f \simeq h$. ∎
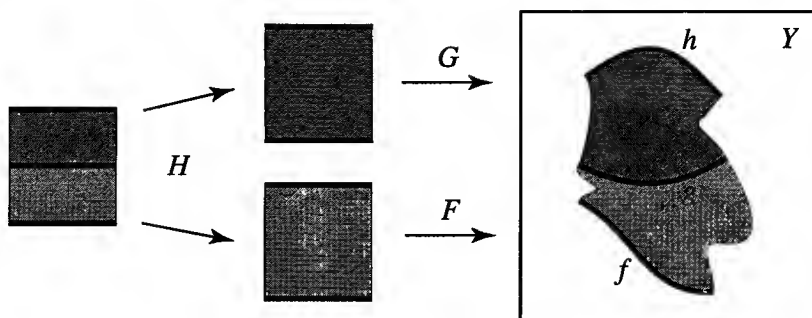
FIGURE 9.3: Forming H from F and G.

Since an equivalence relation on a set yields a partition of the set into equivalence classes, the set $C(X, Y)$ of all continuous functions $f : X \to Y$ is partitioned into equivalence classes under the relation $\simeq$. Consequently, we have the following definition:

DEFINITION 9.3. *Let $C(X, Y)$ denote the set of all continuous functions $f : X \to Y$. Define the **homotopy classes** in $C(X, Y)$ to be the equivalence classes under the relation $\simeq$. The homotopy class containing a function f is denoted $[f]$.*

EXAMPLE 9.3. Every continuous map f from $\mathbb{R}$ to $\mathbb{R}$ is homotopic to the trivial map $g_0 : \mathbb{R} \to \mathbb{R}$ given by $g_0(x) = 0$. The function $F(x, t) = (1 - t)f(x)$ is a homotopy between f and g_0. It follows that the collection $C(\mathbb{R}, \mathbb{R})$ consists of the single homotopy class $[g_0]$.

Exercises for Section 9.1

9.1. Show that the identity map from the disk D in the plane to itself is homotopic to the map that takes D to the origin.

9.2. Show that if $f_1, f_2 : X \to Y$ are homotopic and $g_1, g_2 : Y \to Z$ are homotopic, then $g_2 \circ f_2$ is homotopic to $g_1 \circ f_1$.

9.3. Let f and g be paths in $\mathbb{R}$. Show that f is homotopic to g.

9.4. Let f and g be paths in $\mathbb{R}^2 - \{O\}$. Show that f is homotopic to g. (Hint: Show that every path is homotopic to the constant path that sends the entire interval to the path's starting point. Then show that two constant paths are homotopic using the fact that $\mathbb{R}^2 - \{O\}$ is path connected.)

9.5. Let $f : I \to Y$ and $g : I \to Y$ be paths in Y such that $f(0) = g(0) = y_0$ and $f(1) = g(1) = y_1$. (Therefore f and g have the same initial point and the same endpoint.) A **path homotopy** from f to g defines a continuous function $F : I \times I \to Y$ such that $F(x, 0) = f(x)$, $F(x, 1) = g(x)$, $F(0, t) = y_0$, and $F(1, t) = y_1$.

   (a) Prove that path homotopy defines an equivalence relation on the set of all paths in Y that share the same initial point and the same endpoint.

(b) Show that every path $p : I \to \mathbb{R}$ beginning at 0 and ending at n is path homotopic to the direct path in $\mathbb{R}$ from 0 to n, that is, to the path $p_n : I \to \mathbb{R}$ given by $p_n(x) = nx$.

9.6. Show that if X is a topological space, and D is the disk in the plane, then there is only one homotopy class of continuous functions from X to D.

9.7. Consider the following definition:

> **DEFINITION 9.4.** *A topological space X is said to be* **contractible** *if the identity function on X is homotopic to a constant function.*

(a) Prove that contractibility is a topological invariant. That is, prove that if X and Y are homeomorphic, then X is contractible if and only if Y is contractible.

(b) Prove that $\mathbb{R}^n$ is contractible.

(c) Let X be a contractible space. Prove that X is path connected.

9.8. Let Y have n path components. Prove that there are n homotopy classes of paths in $C(I, Y)$.

9.2   Circle Functions, Degree, and Retractions

Our primary focus in this section is on continuous functions $f : S^1 \to S^1$. Such functions are called **circle functions**. Even though we focus on the circle S^1, the results presented here carry over to every space X that is homeomorphic to the circle. We define the degree of a circle function, and then use degree to prove the Two-Dimensional No Retraction Theorem, which states that there is no continuous function that takes the disk D in the plane to its circle boundary S^1 while fixing S^1.

To represent the points on the circle, we use the variable θ, where θ is the usual angular measure taken from the positive x-axis in the plane. (See Figure 9.4.) We assume two values θ_1 and θ_2 represent the same point on the circle if they differ by an integer multiple of 2π.
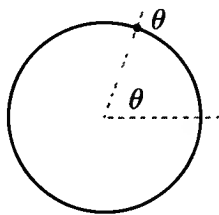
FIGURE 9.4: We represent the points on the circle with the variable θ.

To visualize the behavior of circle functions, we can display their graphs. We think of the graph of a function $f : S^1 \to S^1$ as the set

$$\{(x, y) \in S^1 \times S^1 \mid y = f(x)\}.$$

The graph is a subset of the torus, and if we consider the torus as a square $[0, 2\pi] \times [0, 2\pi]$ with opposite edges identified, then we can exhibit the graphs of circle functions as subsets of $[0, 2\pi] \times [0, 2\pi]$ (with it understood that the opposite edges are identified). For example, in Figure 9.5 we display the graphs of $f(\theta) = \frac{3\pi}{2}$, $g(\theta) = \theta + \frac{\pi}{2}$, $h(\theta) = 2\theta$, and $k(\theta) = -\theta$.
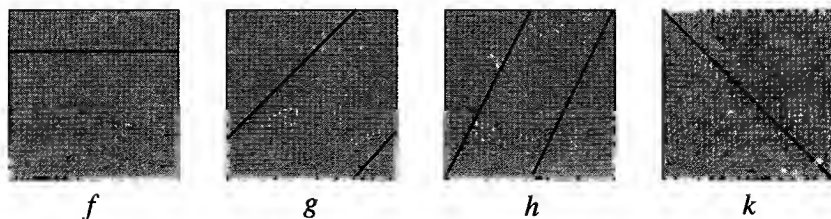
FIGURE 9.5: Graphs of the circle functions f, g, h, and k.

Our goal is to make precise the notion of degree, a measure of how many times a circle function $f : S^1 \to S^1$ wraps the circle around itself. We would like this measure to be equal to 1 for the identity function on the circle, equal to -1 for the function $f(\theta) = -\theta$, and—more generally—equal to n for the function $c_n(\theta) = n\theta$. (See Figure 9.6.)
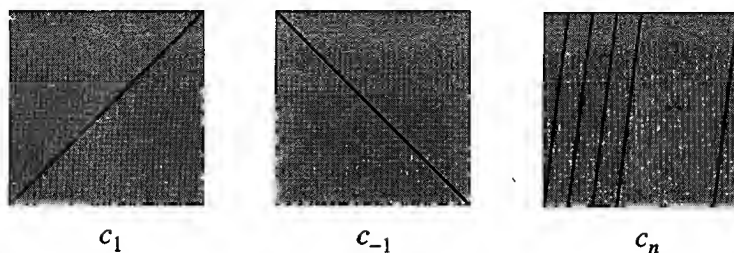
FIGURE 9.6: The functions c_n wrap n times around the circle.

A circle map, not quite as nice as $c_n(\theta) = n\theta$, might oscillate as it wraps around the circle. (See Figure 9.7.) Or it might wind a few times around the circle and then backtrack a few times. It could also pause for a while as it wraps around the circle. Or it might possess a variety of these and other types of complicated behaviors.
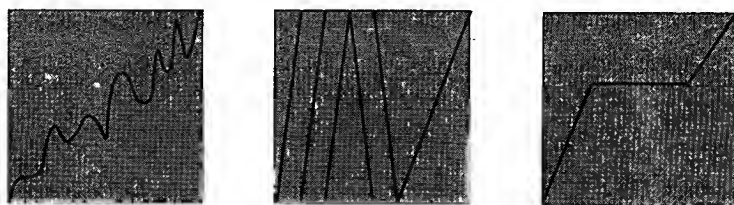
FIGURE 9.7: Oscillations, backtracks, and pauses are some of the complicated behaviors we may observe in circle maps.

The goal behind defining the degree of a circle map is to deform away all of the complicated behavior via homotopy. Fortunately, we can do this as a result of the following theorem:

THEOREM 9.5. *For each circle function, $f : S^1 \to S^1$, there exists a unique $n \in \mathbb{Z}$ such that f is homotopic to $c_n(\theta) = n\theta$.*

DEFINITION 9.6. *The unique n associated to* $f : S^1 \to S^1$ *in Theorem 9.5 is defined to be the **degree of** f and is denoted **deg(f)**.*

We prove Theorem 9.5 in Section 9.6. From the definition, we obtain the following theorem, which indicates that degree determines the homotopy classes of circle functions:

THEOREM 9.7. *Two circle functions* $f, g : S^1 \to S^1$ *are homotopic if and only if* $deg(f) = deg(g)$.

> *Proof.* Since homotopy is an equivalence relation, the distinct homotopy classes of circle maps are disjoint sets. Theorem 9.5 implies that they are given by $\{[c_n(\theta)] \mid n \in \mathbb{Z}\}$. Hence $f \simeq g$ if and only if $[f] = [g] = [c_n(\theta)]$. This occurs exactly when $\deg(f) = \deg(g) = n$.   ∎

The following corollary of Theorem 9.7 is used in our application of degree to a heartbeat model in the next section:

COROLLARY 9.8. *Let* $F : S^1 \times [a, b] \to S^1$ *be continuous. If for each* $r \in [a, b]$ *we let* $F_r : S^1 \to S^1$ *be the circle function defined by* $F_r(\theta) = F(\theta, r)$, *then the degree of* F_r *is independent of* r.

> *Proof.* See Exercise 9.10.   ∎

Corollary 9.8 indicates that if we have a continuous function F mapping an annulus to the circle, then, as illustrated in Figure 9.8, if we restrict F to a circle of radius r, concentric with the annulus, then the degree of the resulting circle function F_r does not depend on r.
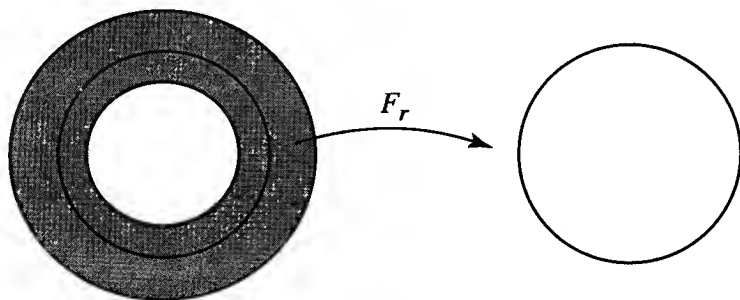
FIGURE 9.8: The degree of the restriction F_r does not depend on the radius r.

The next theorem is an extension theorem concerning circle maps with degree 0; it will be of use to us throughout the chapter.

THEOREM 9.9. *A circle function* $f : S^1 \to S^1$ *has degree 0 if and only if* f *extends to a continuous function on the disk D (that is, if and only if there exists a continuous function* $F : D \to S^1$ *such that* $F(x) = f(x)$ *for all* $x \in S^1$).

Proof. Throughout the proof, we represent the points in the disk with polar coordinates (r, θ).

First, assume that f extends to a continuous function $F : D \to S^1$. Define a function $G : S^1 \times I \to S^1$ by $G(\theta, t) = F(t, \theta)$. Since F is continuous, so is G. Therefore G is a homotopy between the circle functions $G|_{S^1 \times \{0\}}$ and $G|_{S^1 \times \{1\}}$, implying that these functions have the same degree. The function $G|_{S^1 \times \{0\}}$ is given by

$$G|_{S^1 \times \{0\}}(\theta) = G(\theta, 0) = F(0, \theta) = F(0, 0).$$

It follows that $G|_{S^1 \times \{0\}}$ is a constant function and therefore has degree 0. Thus, $G|_{S^1 \times \{1\}}$ also has degree 0. It is straightforward to see that $G|_{S^1 \times \{1\}}$ is equal to f, implying that f has degree 0, as desired.

Now assume that f has degree 0. Therefore there exists a homotopy $G : S^1 \times I \to S^1$ such that $G(\theta, 0) = c_0(\theta)$ and $G(\theta, 1) = f(\theta)$, where $c_0 : S^1 \to S^1$ is the constant function sending each point $\theta \in S^1$ to the point $0 \in S^1$. Define $F : D \to S^1$ by $F(r, \theta) = G(\theta, r)$. Since $G(\theta, r)$ is constant in θ when $r = 0$, it follows that F is well defined at $r = 0$ and therefore is defined as a function on D. Furthermore, G being continuous implies that F is as well. Finally, $F(1, \theta) = G(\theta, 1) = f(\theta)$, implying that F is an extension of f.   ∎

The following result is another useful fact that can be determined from the degree of a circle function:

THEOREM 9.10. *If a circle function $f : S^1 \to S^1$ has a nonzero degree, then f is surjective.*

Proof. See Exercise 9.11.   ∎

We now introduce the concept of a retraction, which is a continuous function that maps a space onto a subspace while fixing the subspace.

DEFINITION 9.11. *Let X be a topological space and A be a subset of X. A **retraction** from X onto A is a continuous function $r : X \to A$ such that $r(a) = a$ for every $a \in A$. If there exists a retraction $r : X \to A$, then A is said to be a **retract** of X. (See Figure 9.9.)*
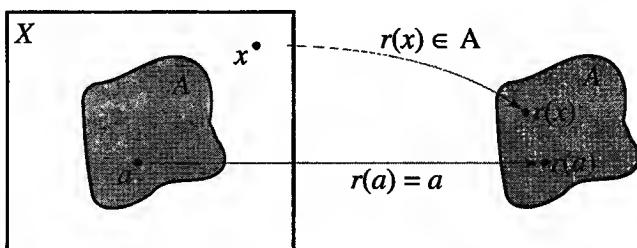
FIGURE 9.9: The subset A is a retract of X.

EXAMPLE 9.4. If A is a one-point subset $\{a\}$ of a topological space X, then A is a retract of X since the function mapping each x in X to a is a retraction.

EXAMPLE 9.5. The circle S^1 is a retract of $\mathbb{R}^2 - \{O\}$. Using (r, θ) polar coordinates, the function $f : \mathbb{R}^2 - \{O\} \to S^1$, defined by $f(r, \theta) = (1, \theta)$, is a retraction.

EXAMPLE 9.6. Consider the annular region A in the plane, shown in Figure 9.10. It is the region centered at the origin, running from the circle C_1 of radius 1 to the circle C_2 of radius 2, inclusive. Individually, C_1 and C_2 are retracts of A. However, together $C_1 \cup C_2$ is not a retract of A because the continuous image of a connected space must be connected.
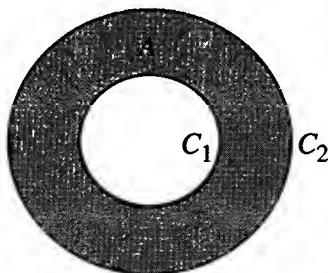
FIGURE 9.10: Individually, C_1 and C_2 are retracts of the annulus, but $C_1 \cup C_2$ is not.

In the previous example, the fact that the circles C_1 and C_2 are retracts of the annulus A is a special case of the next example, which shows that the spaces X and Y can naturally be seen as retracts of the product space $X \times Y$.

EXAMPLE 9.7. Let topological spaces X and Y be given, and consider the product space $X \times Y$. Assume $x^* \in X$. Then the subset $A = \{x^*\} \times Y$ of $X \times Y$ is a retract via the retraction $r : X \times Y \to A$ given by $r(x, y) = (x^*, y)$. Therefore, the product space $X \times Y$ contains subspaces homeomorphic to Y that are retracts (and similarly contains subspaces homeomorphic to X that are retracts).

For example, within the torus there are circle subspaces that are retracts of the torus. (See Figure 9.11.)
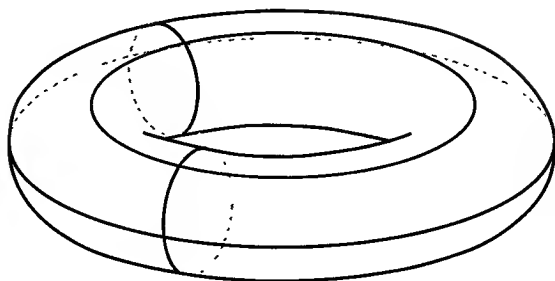
FIGURE 9.11: Circles that are retracts of the torus.

Now, consider the disk D in the plane. We can believe that there is no retraction from D to its circle boundary, just as we might expect that it is not possible to deform the entire skin of a drum to its rim, keeping the points on the rim fixed, but not tearing the skin. (See Figure 9.12.)
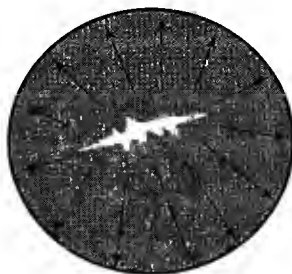
FIGURE 9.12: Deforming the skin of a drum to the rim will tear it, demonstrating that there is no retraction from D onto S^1.

In fact, there is no retraction of the n-ball B^n in $\mathbb{R}^n$ onto its $(n-1)$-sphere boundary, S^{n-1}. This general result is known as the No Retraction Theorem. Here we prove the one- and two-dimensional versions of the No Retraction Theorem. With the use of appropriate tools from algebraic topology, the general result is straightforward.

THEOREM 9.12. *The One-Dimensional No Retraction Theorem.*
There is no retraction from $B^1 = [-1, 1]$ onto its sphere boundary, $S^0 = \{-1, 1\}$.

> ***Proof.*** That there is no retraction from $[-1, 1]$ onto $\{-1, 1\}$ is immediate since a retraction from $[-1, 1]$ onto $\{-1, 1\}$ would be a continuous function from a connected space onto a disconnected one, an impossibility. ∎

THEOREM 9.13. *The Two-Dimensional No Retraction Theorem.*
There is no retraction from the disk D onto its circle boundary, S^1.

Proof. We prove this by contradiction. Thus assume that there exists a continuous function $F : D \to S^1$ such that $F(\theta) = \theta$ for all $\theta \in S^1$. Such a function F is a continuous extension of the identity function $id : S^1 \to S^1$, defined by $id(\theta) = \theta$. The function id has degree 1, and therefore Theorem 9.9 implies that id cannot be extended to a continuous function defined on D, a contradiction. Thus there is no retraction from D onto S^1.   ∎

EXAMPLE 9.8. Theorem 9.13 implies that there is no retraction from $\mathbb{R}^2$ onto S^1, because such a retraction, when restricted to D, would then provide a retraction from D onto S^1.

We revisit the No Retraction Theorem in Chapter 10, where we show that it is equivalent to the Brouwer Fixed Point Theorem.

Next we prove a general existence theorem about retractions. Even though there is no retraction from the disk D onto S^1, if $A \subset D$ is an arc, then no matter how much A winds around in the disk, A is a retract of the disk. (See Figure 9.13.) The proof of this result uses the Tietze Extension Theorem, which was established in a set of supplementary exercises in Section 7.3.
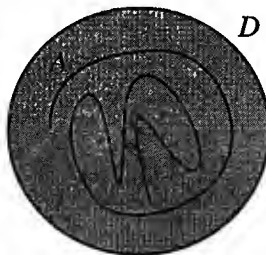
FIGURE 9.13: Every arc A in D is a retract of D.

THEOREM 9.14. *Let D be the disk, and let $A \subset D$ be an arc. Then A is a retract of D.*

Proof. Since A is an arc, there is an embedding $f : [-1, 1] \to D$ mapping onto A. The interval $[-1, 1]$ is compact, and therefore so is A, implying that A is a closed subset of the Hausdorff space D. By the Tietze Extension Theorem, the function $f^{-1} : A \to [-1, 1]$ has a continuous extension $g : D \to [-1, 1]$. The function $f \circ g : D \to A$ is a retraction from D onto A.   ∎

Exercises for Section 9.2

9.9. Determine the degree of each of the following circle functions:
  (a) The antipodal map, $A : S^1 \to S^1$, $A(\theta) = \theta + \pi$, that maps each point on the circle to the point opposite it through the center.

(b) A function that wraps the circle around itself once in one direction and then back around once in the other direction,

$$f : S^1 \to S^1, \, f(\theta) = \begin{cases} 2\theta & \text{for } \theta \in [0, \pi], \\ -2\theta & \text{for } \theta \in [\pi, 2\pi]. \end{cases}$$

9.10. **Prove Corollary 9.8:** Let $F : S^1 \times [a, b] \to S^1$ be continuous. If for each $r \in [a, b]$ we let $F_r : S^1 \to S^1$ be the circle function defined by $F_r(\theta) = F(\theta, r)$, then the degree of F_r is independent of r.

9.11. **Prove Theorem 9.10:** If a circle function $f : S^1 \to S^1$ has a nonzero degree, then f is surjective. (Hint: Prove the contrapositive.)

9.12. Let X be a connected Hausdorff space and B be a two-point subset of X. Prove that B is not a retract of X.

9.13. Let X be a Hausdorff space and A be a retract of X. Prove that A is a closed subset of X.

9.14. Illustrate how the infinite spiral S in Figure 9.14 is a retract of the plane.
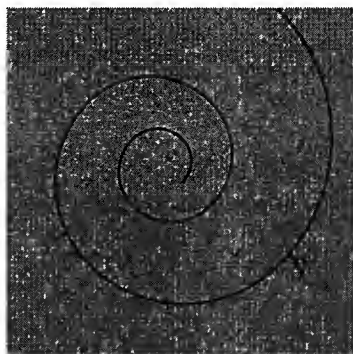
FIGURE 9.14: The spiral S is a retract of the plane.

9.3 *An Application to a Heartbeat Model*

In this section we examine a model for stimulating a beating heart. We use degree theory to draw some interesting conclusions about how the heart responds to stimuli of varied strengths, applied at different times during the beat cycle. This model is adapted from a paper, "Sudden Cardiac Death: A Problem in Topology," by Arthur Winfree (1942–2002), a theoretical biologist who was well regarded for his work on mathematical modeling of biological systems.

We model the time in the beat cycle by a variable θ on a circle (see Figure 9.15), and we assume that the heart beats at $\theta = 0$. We can think of a clock hand rotating around the circle, and each time it passes by $\theta = 0$, the heart beats. Notice that this means we are assuming a beat interval of 2π, for our convenience.

A stimulus is applied to the heart during its beat cycle and we wish to examine the response of the heart to the stimulus. The response is measured as the time it takes the heart to beat again after the stimulus is applied; this variable is known as the **latency** and is denoted L.
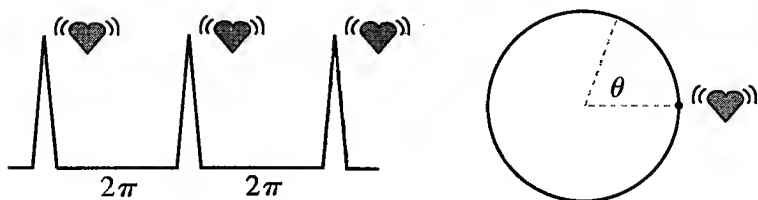
FIGURE 9.15: Time in the beat cycle is modeled by a circular variable θ.

The time in the beat cycle at which the stimulus is applied is called the **coupling interval**; we denote it by c. We consider stimuli of varying strengths s from weak ($s = w$) to strong ($s = g$) and examine $L(c, s)$, the latency as a function of coupling interval and stimulus strength. Thus, we have a function $L : S^1 \times [w, g] \rightarrow S^1$.

The heart's response to a weak stimulus is called **weak rescheduling**, and its response to a strong stimulus is called **strong rescheduling**. We make the following reasonable assumptions about weak and strong rescheduling:

(i) With weak rescheduling, the stimulus is too weak to result in any appreciable change in the beat timing, and therefore the latency is the time remaining in the beat cycle after the stimulus.

(ii) With strong rescheduling, the stimulus is strong enough so that the heart loses memory of the past, resulting in a latency that is independent of the coupling interval.

These assumptions translate to the following assumed properties of the latency L, as illustrated in Figure 9.16:

(i) $L(c, w) = 2\pi - c$ for all $c \in S^1$.

(ii) There exists $\theta^* \in S^1$ such that $L(c, g) = \theta^*$ for all $c \in S^1$.
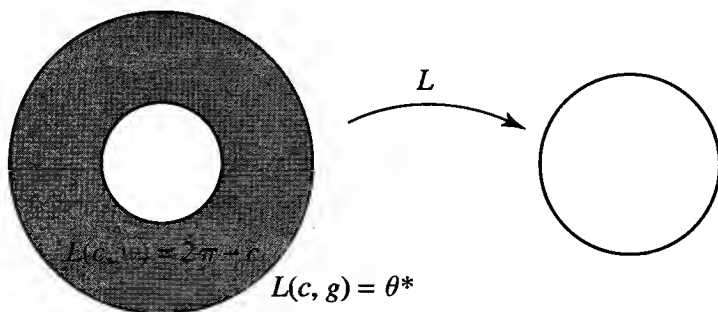
FIGURE 9.16: Latency $L(c, s)$ defined over the domain $S^1 \times [w, g]$.

If we restrict L to a particular stimulus strength s^*, then the result is a circle function $L|_{S^1 \times \{s^*\}}$. If $L|_{S^1 \times \{s^*\}}$ is continuous, then it has a well-defined degree.

The function $L|_{S^1 \times \{w\}}$ has degree -1 since it wraps $S^1 \times \{w\}$ around S^1 once in the clockwise direction. Furthermore, $L|_{S^1 \times \{g\}}$ has degree 0 since it is a constant function. But Corollary 9.8 implies that if L is continuous, then the degree of $L|_{S^1 \times \{s^*\}}$ is independent of s^*. Hence, $L : S^1 \times [w, g] \to S^1$ is not continuous; that is, given our assumptions, latency does not depend continuously on stimulus timing and strength over a strength interval varying from weak to strong.

Thus, there are discontinuities in the heart's response to varying stimulus timing and strength. If the heart receives a stimulus at the appropriate time in the beat cycle, that stimulus can disrupt the beat pattern of the heart in an unpredictable way.

Now, let us examine $L : S^1 \times [w, g] \to S^1$ further to extract possible additional topological consequences of our assumptions. Since L is not continuous, there is a set of points $P \subset S^1 \times [w, g]$ where L fails to be continuous. We consider the situation where this set of discontinuities lies in the interior of a disk K in $S^1 \times [w, g]$. (See Figure 9.17.) For example, this occurs if there is a single point of discontinuity of L and that point is in the interior of $S^1 \times [w, g]$. We claim that on the circle that is the boundary of K, the latency function L takes on all values of latencies from 0 to 2π.
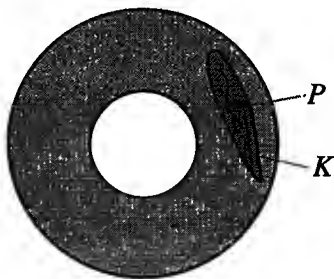
FIGURE 9.17: The set P, of discontinuities of L, lies in K.

To prove this claim, we make some changes to L and its domain. First, note that if A is the set obtained by removing the interior of K from $S^1 \times [w, g]$, then $L|_A$, by assumption, is continuous. To see the next stage of our construction, we pictorially flip A so that the inner circle $s = w$ is on the outside and the outer circle $s = g$ is on the inside. (See Figure 9.18.)

Next, we fill in the inner-diameter hole to obtain a new domain A'. Furthermore, we extend L to a continuous function L' by setting L' equal to the constant θ^* on the filled-in portion of A'. (See Figure 9.19.) Since L' was already equal to θ^* on the inner boundary of A, it follows that L' is a continuous function. (Note that if $L|_{S^1 \times \{g\}}$ was not necessarily constant, but had degree 0, then by Theorem 9.9, we could still extend $L|_A$ to a continuous function L' on A'.)
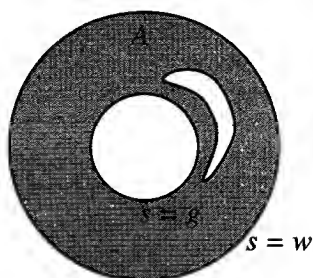
FIGURE 9.18: The domain A turned inside out.

FIGURE 9.19: Extending L on A to L' defined on A'.

The domain A' is topologically an annulus. Therefore we can apply Corollary 9.8 to conclude that the degree of $L'|_{\partial K}$ equals the degree of $L'|_{s=w}$. As we discussed previously, the degree of $L'|_{s=w}$ is equal to -1. Theorem 9.10 implies that if we consider $L'|_{\partial K}$ as a function from the circle to itself, then it is surjective. Therefore, as claimed, on the circle that is the boundary of K, the function L takes on all values of latencies from 0 to 2π.

Thus, we see that a quite complicated distribution of responses can arise near the set of discontinuities of L. In particular, if L has a single point of discontinuity p, and that point is interior to $S^1 \times [w, g]$, then on every circle surrounding p, the function L takes on all possible latency values. In such a situation, it follows that with very small variations of stimulus strength and timing away from p, the whole spectrum of latency responses can be experienced. We cannot predict how long it will take in the beat cycle before the heart beats again. Anything can happen.

The topological properties of the model presented here provide possible explanations for the cause of fibrillation, an often fatal disorganization of the heart's contraction pattern. George Mines, a McGill University experimental researcher of fibrillation, met an unfortunate death at the age of 28 in 1914 when he chose to be his own experimental subject. The laboratory janitor discovered Mines, unconscious, with the experimental apparatus attached to his chest, the heartbeat monitor recording his failing heartbeat. Mines died later that evening. In his final paper, Mines had hypothesized that fibrillation can be triggered by an untimely electrical impulse. The topological properties of the model just presented, as well as the fatal results of Mines's experiment, certainly support his conjecture.

Exercises for Section 9.3

9.15. This exercise provides an alternative approach to showing that the latency function $L : S^1 \times [w, g] \to S^1$ is not continuous. Assume that L is continuous, and from it construct a continuous function on the disk in the plane, $f : D \to S^1$, such that the restriction to the boundary circle, $f|_{S^1} : S^1 \to S^1$, has degree -1. Use Theorem 9.9 to derive a contradiction.

9.4   The Fundamental Theorem of Algebra

We have previously seen the Intermediate Value Theorem and the Extreme Value Theorem, two important applications of topology in analysis. In this section, we show how degree can be used to prove an important basic theorem in algebra, the Fundamental Theorem of Algebra.

The Fundamental Theorem of Algebra asserts that every polynomial equation

$$a_n z^n + a_{n-1} z^{n-1} + \ldots + a_1 z + a_0 = 0 \tag{9.1}$$

has a solution within the complex plane, $\mathbb{C}$. It is surprising that we can prove such a fundamental algebraic result using the tools of topology.

First, note that since we can divide Equation 9.1 through by a_n, we only need to consider equations of the form

$$z^n + a_{n-1} z^{n-1} + \ldots + a_1 z + a_0 = 0.$$

This equation has a solution in $\mathbb{C}$ if and only if for each positive real number c, the equation

$$z^n + \frac{a_{n-1}}{c} z^{n-1} + \ldots + \frac{a_1}{c^{n-1}} z + \frac{a_0}{c^n} = 0$$

has a solution in $\mathbb{C}$. (See Exercise 9.16.) By choosing a c large enough, we can ensure that

$$\left| \frac{a_{n-1}}{c} \right| + \ldots + \left| \frac{a_1}{c^{n-1}} \right| + \left| \frac{a_0}{c^n} \right| < 1.$$

Therefore, to establish the Fundamental Theorem of Algebra, we only need to consider equations of the form

$$z^n + b_{n-1} z^{n-1} + \ldots + b_1 z + b_0 = 0,$$

with $|b_{n-1}| + \ldots + |b_0| < 1$. These equations are the subject of the following theorem:

THEOREM 9.15. *If $|b_{n-1}| + \ldots + |b_0| < 1$, then the equation*

$$z^n + b_{n-1} z^{n-1} + \ldots + b_1 z + b_0 = 0$$

has a solution in $\mathbb{C}$.

Proof. We establish this result by contradiction. Assume that

$$z^n + b_{n-1}z^{n-1} + \ldots + b_1z + b_0 = 0 \tag{9.2}$$

has no solution within the complex plane. Let D be the disk and S^1 be the circle, here considered as subsets of the complex plane. Since Equation 9.2 has no solution, we can define a continuous function $F : D \to \mathbb{C} - \{0\}$ by

$$F(z) = z^n + b_{n-1}z^{n-1} + \ldots + b_1z + b_0.$$

The function F is an extension of the function $f : S^1 \to \mathbb{C} - \{0\}$, where $f(z) = F(z)$.

We define new continuous functions $g : S^1 \to S^1$ and $G : D \to S^1$ by $g(z) = \frac{f(z)}{|f(z)|}$ and $G(z) = \frac{F(z)}{|F(z)|}$. The function G is an extension of g; therefore Theorem 9.9 implies that g has degree 0.

We claim that g is homotopic to $c_n(z) = z^n$. The function c_n is the function that we previously expressed as $c_n(\theta) = n\theta$; here it is expressed as a function of the complex variable z, rather than as a function of the angle θ. Given the claim, we have a contradiction since g has degree 0 and c_n has degree $n \neq 0$.

To prove the claim, define $H : S^1 \times [0, 1] \to S^1$ by

$$H(z, t) = \frac{(z^n + t(b_{n-1}z^{n-1} + \ldots + b_1z + b_0))}{|z^n + t(b_{n-1}z^{n-1} + \ldots + b_1z + b_0)|}.$$

To establish that H is well defined, we need to show that for all $(z, t) \in S^1 \times [0, 1]$,

$$|z^n + t(b_{n-1}z^{n-1} + \ldots + b_1z + b_0)| > 0. \tag{9.3}$$

Since $t \in [0, 1]$, $|b_{n-1}| + \ldots + |b_0| < 1$, and $|z^j| = 1$ for all $z \in S^1$ and $j \in \mathbb{Z}_+$, we have

$$
\begin{aligned}
|z^n + t(b_{n-1}z^{n-1} + \ldots + b_1z + b_0)| &\geq |z^n| - t|b_{n-1}z^{n-1} + \ldots + b_1z + b_0| \\
&\geq 1 - t(|b_{n-1}z^{n-1}| + \ldots + |b_1z| + |b_0|) \\
&= 1 - t(|b_{n-1}| + \ldots + |b_0|) \\
&> 1 - t \\
&\geq 0.
\end{aligned}
$$

From the expression for $H(z, t)$ and Inequality 9.3, it follows that H is a well-defined continuous function mapping $S^1 \times [0, 1]$ to S^1. Furthermore, $H(z, 1) = \frac{f(z)}{|f(z)|} = g(z)$ and $H(z, 0) = \frac{z^n}{|z^n|} = z^n$, with the latter equality holding because $|z^n| = 1$ for all $z \in S^1$. Therefore, H is a homotopy between g and c_n, and the proofs of the claim and the theorem are complete. ∎

Exercises for Section 9.4

9.16. Show that the equation

$$z^n + a_{n-1}z^{n-1} + \ldots + a_1 z + a_0 = 0$$

has a solution in $\mathbb{C}$ if and only if, for each positive real number c, the equation

$$z^n + \frac{a_{n-1}}{c}z^{n-1} + \ldots + \frac{a_1}{c^{n-1}}z + \frac{a_0}{c^n} = 0$$

has a solution in $\mathbb{C}$.

9.17. Provide justifications for each step in the following portion from the proof of Theorem 9.15:

$$
\begin{aligned}
|z^n + t(b_{n-1}z^{n-1} + \ldots + b_1 z + b_0)| &\geq |z^n| - t|b_{n-1}z^{n-1} + \ldots + b_1 z + b_0| \\
&\geq 1 - t(|b_{n-1}z^{n-1}| + \ldots + |b_1 z| + |b_0|) \\
&= 1 - t(|b_{n-1}| + \ldots + |b_0|) \\
&> 1 - t \\
&\geq 0.
\end{aligned}
$$

9.18. (a) Prove that if c is such that

$$\left|\frac{a_{n-1}}{c}\right| + \ldots + \left|\frac{a_1}{c^{n-1}}\right| + \left|\frac{a_0}{c^n}\right| < 1,$$

then all of the solutions to

$$z^n + a_{n-1}z^{n-1} + \ldots + a_1 z + a_0 = 0$$

satisfy $|z| < c$. (Hint: Show that if $|z| \geq c$, then

$$\left|\frac{a_{n-1}}{z} + \ldots + \frac{a_1}{z^{n-1}} + \frac{a_0}{z^n}\right| < 1,$$

and, from that, prove that no such z can be a solution to the equation.)

(b) Show that all solutions to the equation

$$z^n + z^{n-1} + \ldots + z + 1 = 0$$

lie in the open disk of radius 2 centered at the origin in the complex plane.

9.5 More on Distinguishing Topological Spaces

In this section we extend results from Section 6.2 on distinguishing between topological spaces. In Section 6.2, we used the concept of connectedness to prove that a variety of spaces are not homeomorphic to each other. Here, we use the concept of simple connectedness in a similar way.

DEFINITION 9.16. *A path connected topological space X is called **simply connected** if every continuous function $f : S^1 \to X$ is homotopic to a constant function.*

If we think of a continuous function $f : S^1 \to X$ as a loop in X, a path connected topological space is simply connected if every loop in X can be deformed to a point in X. (See Figure 9.20.)
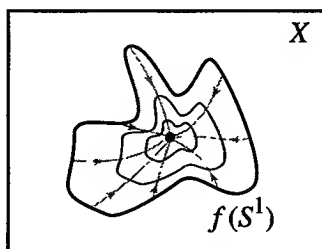
FIGURE 9.20: The space X is simply connected if every loop in X can be deformed to a point.

EXAMPLE 9.9. It is not difficult to see that the plane $\mathbb{R}^2$ is simply connected. Every continuous function $f : S^1 \to \mathbb{R}^2$ is homotopic to a constant function sending S^1 to the origin. Such a homotopy is accomplished by using the straight-line homotopy $F : S^1 \times I \to \mathbb{R}^2$, defined by $F(x, t) = (1 - t)f(x)$. On the other hand, if we remove the origin from the plane, then the resulting space $\mathbb{R}^2 - \{O\}$ is not simply connected. A loop wrapping once around the origin cannot be deformed to a point in the plane. (See Figure 9.21.) We prove this in Theorem 9.19.
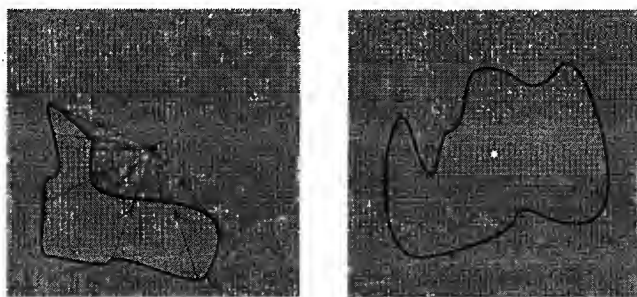
FIGURE 9.21: The plane $\mathbb{R}^2$ is simply connected, but $\mathbb{R}^2 - \{O\}$ is not.

The following two theorems concerning simple connectedness are straightforward. We leave their proofs to the exercises.

THEOREM 9.17. *Simple connectedness is a topological property. That is, if X is homeomorphic to Y, then X is simply connected if and only if Y is simply connected.*

Proof. See Exercise 9.20.   ∎

THEOREM 9.18. *If X is simply connected and A is a retract of X, then A is simply connected.*

    Proof. See Exercise 9.21. ∎

    In Section 6.2, when we showed that the line is not homeomorphic to the plane, we did so by showing that removing a point from the line results in a disconnected space, while removing a point from the plane does not. The same idea works for us here in showing that the plane is not homeomorphic to $\mathbb{R}^3$. Specifically, we prove that removing a point from the plane results in a space that is not simply connected, while removing a point from $\mathbb{R}^3$ results in a space that is simply connected. These results should be intuitively clear. We work out the details next.

THEOREM 9.19. *The space $\mathbb{R}^2 - \{O\}$ is not simply connected.*

    Proof. We show that there is a continuous function $f : S^1 \to \mathbb{R}^2 - \{O\}$ that is not homotopic to a constant function. To do this, let f be the embedding of S^1 onto itself as a subspace of $\mathbb{R}^2 - \{O\}$. We claim that there is no homotopy from f to a constant function.

    Suppose that there is a homotopy $F : S^1 \times I \to \mathbb{R}^2 - \{O\}$ from f to a constant function. Then, if we define $G : S^1 \times I \to S^1$ by $G(x,t) = \frac{F(x,t)}{|F(x,t)|}$, it follows that G is a homotopy from the identity function on the circle to a constant function on the circle. But the identity function on the circle has degree 1, and a constant function has degree 0. Therefore, by Theorem 9.7 no such homotopy G exists. Thus we have a contradiction to our assumption that F exists. It follows that f is not homotopic to a constant function, and $\mathbb{R}^2 - \{O\}$ is not simply connected. ∎

    Now, let α be a point in $\mathbb{R}^3$. Assume that $f : S^1 \to \mathbb{R}^3 - \{\alpha\}$ is continuous. It is apparent that there is enough room to deform f to a constant function while avoiding the missing point at α. (See Figure 9.22.) We show how to construct such a deformation in the proof of the next theorem.
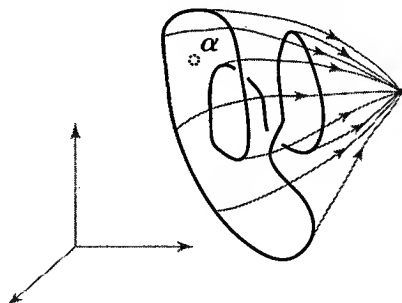
FIGURE 9.22: There is room in $\mathbb{R}^3 - \{\alpha\}$ to deform $f : S^1 \to \mathbb{R}^3 - \{\alpha\}$ to a constant function.

THEOREM 9.20. *For each $\alpha \in \mathbb{R}^3$, the space $\mathbb{R}^3 - \{\alpha\}$ is simply connected.*

Proof. We prove that if $f : S^1 \to \mathbb{R}^3 - \{\alpha\}$ is continuous, then it is homotopic to a constant function. We do so by demonstrating a two-stage homotopy from f to a constant function. However, we only present a description of what takes place under the homotopy, rather than provide a specific definition of it.

To begin, since $f(S^1)$ is compact (and therefore closed) and since $\alpha \notin f(S^1)$, there exists an $\varepsilon > 0$ such that the open ball of radius ε centered at α is disjoint from $f(S^1)$. Using the compactness of $f(S^1)$ again, we can cover $f(S^1)$ by finitely many open balls of radius $\varepsilon/2$. Each such open ball does not contain α.

If we take the preimage under f of each of these open balls, the result is an open cover of S^1. By Corollary 7.28, we can subdivide S^1 with

$$0 = \theta_0 < \theta_1 < \ldots < \theta_m = 2\pi$$

such that for each $i = 0, \ldots, m - 1$, the interval $[\theta_i, \theta_{i+1}]$ is contained in a set in the open cover of S^1. Therefore each set $f([\theta_i, \theta_{i+1}])$ is contained in one of the aforementioned open balls. For each $i = 0, \ldots, m - 1$, pick such an open ball and denote it by B_i. For each $i = 0, \ldots, m$, set $s_i = f(\theta_i)$, and let S_i be the line segment connecting s_i and s_{i+1}. (Note that it is possible that the segment S_i is a single point.) Each S_i lies in the open ball B_i.

In the first stage of the homotopy, we deform each $f([\theta_i, \theta_{i+1}])$ within B_i to S_i. (See Figure 9.23.) This deformation can be done so that each point s_i remains fixed throughout the deformation. Since B_i does not contain α, this deformation takes place in $\mathbb{R}^3 - \{\alpha\}$. Thus $f(S^1)$ has been deformed in $\mathbb{R}^3 - \{\alpha\}$ to a closed curve P that is made up entirely of line segments.
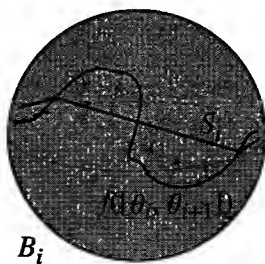
FIGURE 9.23: Deforming $f([\theta_i, \theta_{i+1}])$ to S_i within B_i.

For each $i = 0, \ldots, m - 1$, let A_i be the plane through α, s_i, and s_{i+1} (or the line through α, s_i, and s_{i+1} if they are collinear). We can choose a point $\beta \in \mathbb{R}^3$ such that β does not lie in any of the sets A_i. Note that α does not lie in any of the triangles T_i with vertices β, s_i, and s_{i+1}. (By triangle, here, we include the possibility that $s_i = s_{i+1}$.)

Now, in the second stage of the homotopy, we deform each line segment S_i within T_i to the point β, and the deformations are done so that they all glue together, resulting in a continuous deformation of the closed curve P to the point β. (See Figure 9.24.) This second stage of the homotopy also takes place in $\mathbb{R}^3 - \{\alpha\}$ since α does not lie in any of the sets T_i.
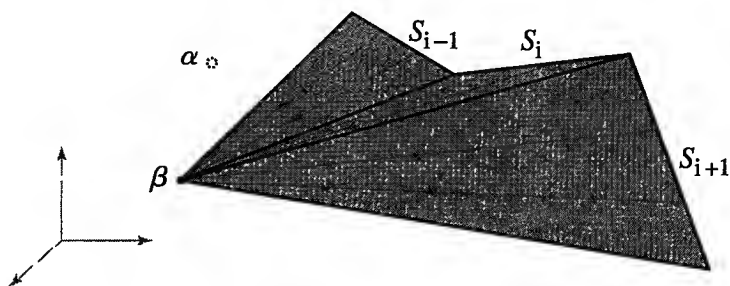
FIGURE 9.24: Deforming each S_i to β.

When we combine the two stages of the homotopy, the result is a homotopy from f to a constant function, and it follows that $\mathbb{R}^3 - \{\alpha\}$ is simply connected for all $\alpha \in \mathbb{R}^3$.   ∎

Now, via Theorems 9.17, 9.19, and 9.20 we can distinguish the plane from 3-space.

THEOREM 9.21. *The plane $\mathbb{R}^2$ is not homeomorphic to 3-space, $\mathbb{R}^3$.*

Proof. Suppose that there is a homeomorphism $f : \mathbb{R}^2 \to \mathbb{R}^3$. Since $\mathbb{R}^2 - \{O\}$ is not simply connected, it would follow that $\mathbb{R}^3 - \{f(O)\}$ is not simply connected as well, a contradiction. Thus no such f exists, and it follows that $\mathbb{R}^2$ is not homeomorphic to $\mathbb{R}^3$.   ∎

The approach we used in proving that $\mathbb{R}^3 - \{\alpha\}$ is simply connected can also be used to prove that $\mathbb{R}^n - \{\alpha\}$ is simply connected for all $n \geq 3$ and all $\alpha \in \mathbb{R}^n$. Thus it follows that $\mathbb{R}^2$ is not homeomorphic to $\mathbb{R}^n$ for all $n \geq 3$. Furthermore, the function $f : \mathbb{R}^{n+1} - \{O\} \to S^n$, given by $f(x) = \frac{x}{|x|}$, is a retraction. Therefore Theorem 9.18 implies that S^n is simply connected for all $n \geq 2$. The circle, S^1, is not simply connected. (See Exercise 9.22.) Hence, S^1 is not homeomorphic to S^n for all $n \geq 2$.

Exercises for Section 9.5

9.19. Prove that a topological space X is simply connected if and only if every continuous function $f : S^1 \to X$ extends to a continuous function $F : D \to X$ where D is the disk in the plane.

9.20. Prove Theorem 9.17: If X is homeomorphic to Y, then X is simply connected if and only if Y is simply connected.

9.21. **(a)** **Prove Theorem 9.18:** If X is simply connected and A is a retract of X, then A is simply connected.

   (b) Provide an example showing that if A is a retract of X and A is simply connected, then X need not be simply connected.

9.22. Prove that S^1 is not simply connected. (Hint: Use an approach similar to that used in the proof of Theorem 9.19.)

9.23. Show that $X \times Y$ is simply connected if and only if X and Y are simply connected.

9.6   More on Degree

In this section we prove Theorem 9.5, the main result establishing the definition of degree and the invariance of degree under homotopy between circle functions. In our view of S^1 in Sections 9.2 and 9.3, we used real values θ to represent points on the circle, understanding that θ is associated to the point with counterclockwise angle θ from the positive x-axis. In this section, we need to reserve real values for the purpose of representing points on the real line. In order to avoid confusing the two interpretations, we represent points on the circle in their complex exponential format. So $e^{i\theta}$ represents the complex number $\cos(\theta) + \sin(\theta)i$ on the circle, located at angle θ from the x-axis. (See Figure 9.25.)
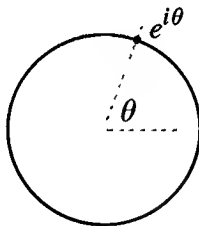
FIGURE 9.25: Representing points on the circle in complex exponential format.

We show that for every circle function $f : S^1 \to S^1$ there is a unique $n \in \mathbb{Z}$ such that f is homotopic to the function $c_n : S^1 \to S^1$, defined by $c_n(e^{i\theta}) = e^{in\theta}$. This c_n is the function obtained by taking the function $f(z) = z^n$, defined on the complex plane, and restricting it to the circle. This function also corresponds exactly to the function $c_n : S^1 \to S^1$ that we previously presented as $c_n(\theta) = n\theta$.

In order to prove the desired result, we associate to each circle function $f : S^1 \to S^1$ a function $f^* : [0, 2\pi] \to \mathbb{R}$ that essentially unwraps f. We then show that a straight-line homotopy exists between f^* and a linear function $c_n^* : [0, 2\pi] \to \mathbb{R}$, defined by $c_n^*(\theta) = n\theta$. From that homotopy we then obtain the desired homotopy between circle functions.

Every circle function $f : S^1 \to S^1$ can be viewed as a continuous function $f : [0, 2\pi] \to S^1$ having $f(0) = f(2\pi)$. Conversely, given a continuous function $f : [0, 2\pi] \to S^1$ satisfying $f(0) = f(2\pi)$, there is a circle function $f : S^1 \to S^1$ naturally associated with it. Thus, in this section we work with

continuous functions $f : [0, 2\pi] \to S^1$ satisfying $f(0) = f(2\pi)$. As with continuous functions $f : S^1 \to S^1$, we also refer to these functions as **circle functions**.

Homotopies between circle functions play an important role in the results we present here. Such a homotopy, $H : [0, 2\pi] \times [0, 1] \to S^1$, must define a circle function for each $t \in [0, 1]$. Thus, we have the following definition:

DEFINITION 9.22. *Circle functions $f, g : [0, 2\pi] \to S^1$ are said to be* **circle homotopic** *if there is a homotopy $H : [0, 2\pi] \times [0, 1] \to S^1$ between f and g such that $H(0, t) = H(2\pi, t)$ for all $t \in [0, 1]$. Such a function H is called a* **circle homotopy.**

A circle homotopy $H : [0, 2\pi] \times [0, 1] \to S^1$ naturally defines a homotopy $H : S^1 \times [0, 1] \to S^1$ between circle functions that have domain S^1. Furthermore, a homotopy $H : S^1 \times [0, 1] \to S^1$ naturally yields a circle homotopy.

Now consider the function $p : \mathbb{R} \to S^1$ defined by $p(\theta) = e^{i\theta}$. The function p maps each interval $[r, r + 2\pi)$ bijectively around the circle and therefore wraps the real line around the circle an infinite number of times. (See Figure 9.26.) It is an example of a type of continuous function known as a covering map.
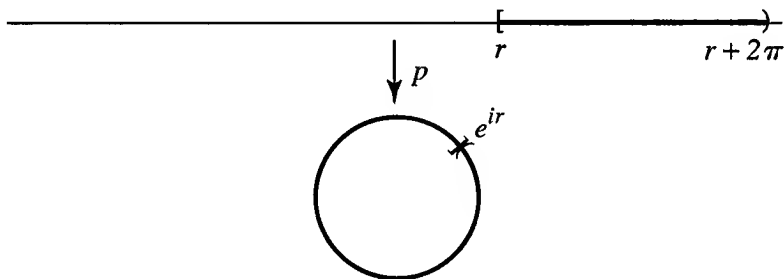
FIGURE 9.26: The function p wraps $[r, r + 2\pi)$ around S^1.

When we restrict p to an open interval (a, b) with $b - a \leq 2\pi$, the resulting function $p|_{(a,b)}$ is a homeomorphism onto its image in S^1. For each $r \in R$, let q_r be the inverse of the homeomorphism $p|_{(r,r+2\pi)}$. The function q_r homeomorphically maps $S^1 - \{e^{ir}\}$, the circle with the point e^{ir} removed, to the interval $(r, r + 2\pi)$, as illustrated in Figure 9.27.

The idea behind unwrapping a circle function $f : [0, 2\pi] \to S^1$ is to associate to it a function $f^* : [0, 2\pi] \to \mathbb{R}$, as follows:

DEFINITION 9.23. *Let $f : X \to S^1$ be continuous. A continuous function $f^* : X \to \mathbb{R}$ is called a* **lifting** *of f if $p \circ f^* = f$. (See Figure 9.28.)*
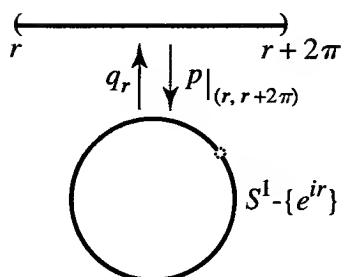
FIGURE 9.27: The functions q_r and $p|_{(r,r+2\pi)}$ are inverse homeomorphisms.
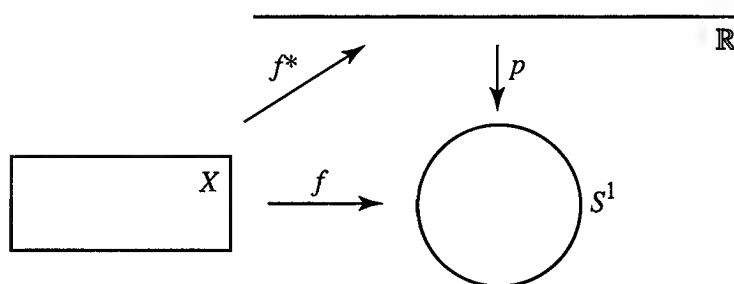
FIGURE 9.28: A lifting f^* of f satisfies $p \circ f^* = f$.

EXAMPLE 9.10. A path wrapping three-quarters of the way around the circle counterclockwise from $1 = e^{i0}$ has a lifting that is a path in $\mathbb{R}$ running from 0 to $3\pi/2$. A path wrapping exactly once around the circle clockwise from $-1 = e^{i\pi}$ has a lifting that is a path in $\mathbb{R}$ running from π to $-\pi$. (See Figure 9.29.)
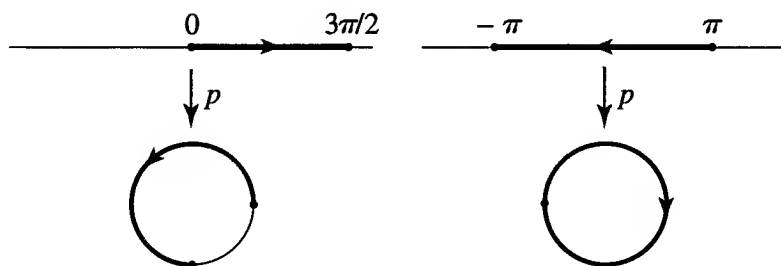
FIGURE 9.29: Liftings of paths on the circle.

The following theorem asserts that liftings exist for a function $f : X \to S^1$, as long as f is not surjective:

THEOREM 9.24. *Let $f : X \to S^1$ be continuous and assume that $f(X)$ is a proper subset of S^1. Furthermore, assume that $x_0 \in X$ and $r_0 \in p^{-1}(f(x_0))$. Then there exists a lifting, $f^* : X \to \mathbb{R}$, of f such that $f^*(x_0) = r_0$.*

Proof. There exists $s \in S^1$ such that $s \notin f(X)$. We can choose $r \in p^{-1}(s)$ such that $r_0 \in (r, r + 2\pi)$. Let q_r be the homeomorphism previously described, and define $f^* : X \to \mathbb{R}$ by $f^*(x) = q_r(f(x))$. It follows that f^* is a lifting of f, and $f^*(x_0) = r_0$. (See Figure 9.30.)
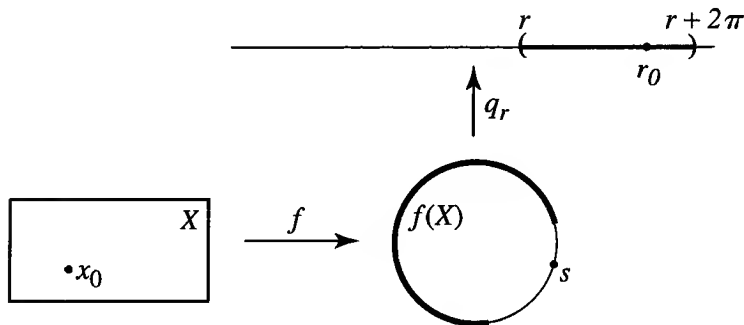
FIGURE 9.30: Lifting a function $f : X \to S^1$ that is not surjective.   ∎

Liftings are not uniquely determined. Since $p(\theta) = p(\theta + 2\pi n)$ for every $\theta \in \mathbb{R}$ and $n \in \mathbb{Z}$, it follows that if f^* is a lifting of $f : X \to S^1$, then for every $n \in \mathbb{Z}$ the function $f_n^* : X \to \mathbb{R}$, defined by $f_n^*(x) = f^*(x) + 2\pi n$, is also a lifting of f. In the case where X is connected, this accounts for all of the possible liftings, as the following theorem indicates:

THEOREM 9.25. *Let $f : X \to S^1$ be continuous and X be connected. If $g, h : X \to \mathbb{R}$ are liftings of f, then there exists $n \in \mathbb{Z}$ such that $g(x) - h(x) = 2\pi n$ for all $x \in X$.*

Proof. Let $x \in X$ be arbitrary. Then $p(g(x)) = f(x) = p(h(x))$, and therefore there exists $n_x \in \mathbb{Z}$ such that $g(x) - h(x) = 2\pi n_x$. Define a function $k : X \to \mathbb{R}$ by $k(x) = g(x) - h(x)$. The image of k is the set $\{2\pi n_x \mid x \in X\}$, and this set must be connected since k is continuous and X is connected. Thus, the set $\{2\pi n_x \mid x \in X\}$ is a connected subspace of $\mathbb{R}$, implying that $n_x = n_y$ for each $x, y \in X$. Hence, there exists $n \in \mathbb{Z}$ such that $g(x) - h(x) = 2\pi n$ for all $x \in X$.   ∎

The next theorem is a straightforward consequence of the previous one. It indicates that if X is connected and two liftings agree at one point in X, then they agree at every point in X.

THEOREM 9.26. *Let $f : X \to S^1$ be continuous, X be connected, and $g, h : X \to \mathbb{R}$ be liftings of f. If there exists $x_0 \in X$ such that $g(x_0) = h(x_0)$, then $g(x) = h(x)$ for all $x \in X$.*

Proof. See Exercise 9.25.   ∎

Now, we are ready to unwrap circle maps. The idea is to show that for every circle map $f : [0, 2\pi] \to S^1$, there exists a lifting $f^* : [0, 2\pi] \to \mathbb{R}$. We establish a slightly more general lifting result that holds for all continuous functions $f : [0, 2\pi] \to S^1$, not just for circle functions.

THEOREM 9.27. *Let $f : [0, 2\pi] \to S^1$ be continuous. Then there exists a function $f^* : [0, 2\pi] \to \mathbb{R}$ that is a lifting of f.*

Proof. Recall that the complex numbers e^{i0} and $e^{i\pi}$ equal 1 and -1, respectively. Let $U = S^1 - \{1\}$ and $V = S^1 - \{-1\}$. The open sets U and V together cover S^1, and therefore $\{f^{-1}(U), f^{-1}(V)\}$ is an open covering of $[0, 2\pi]$. Corollary 7.28 implies that there is a subdivision $0 = \theta_0 < \theta_1 < \ldots < \theta_n = 2\pi$ of $[0, 2\pi]$ such that for all $j = 1, \ldots, n$, the interval $[\theta_{j-1}, \theta_j]$ is contained in either $f^{-1}(U)$ or $f^{-1}(V)$. For each interval $[\theta_{j-1}, \theta_j]$ the restriction $f|_{[\theta_{j-1}, \theta_j]}$ is not surjective, and therefore Theorem 9.24 implies that there is a lifting of $f|_{[\theta_{j-1}, \theta_j]}$. The goal is to build these liftings up, step by step, so that they properly join together to form the desired lifting f^*.

To begin, let $g_1 : [\theta_0, \theta_1] \to \mathbb{R}$ be a lifting of $f|_{[\theta_0, \theta_1]}$. Set f_1^*, the first part of our lifting, equal to g_1. Note that $p(g_1(\theta_1)) = f(\theta_1)$, and therefore $g_1(\theta_1) \in p^{-1}(f(\theta_1))$. Thus, by Theorem 9.24 there exists $g_2 : [\theta_1, \theta_2] \to \mathbb{R}$, a lifting of $f|_{[\theta_1, \theta_2]}$ with $g_2(\theta_1) = g_1(\theta_1)$. By the Pasting Lemma we can paste the functions g_1 and g_2 together to obtain a continuous function $f_2^* : [\theta_0, \theta_2] \to \mathbb{R}$. The function f_2^* is a lifting of $f|_{[\theta_0, \theta_2]}$. We can continue this construction, and the function f_n^* obtained at the nth step is the desired function f^*. ∎

If $f : [0, 2\pi] \to S^1$ is a circle map, then by definition $f(0) = f(2\pi)$. However, it need not be that $f^*(0) = f^*(2\pi)$ for a lifting $f^* : [0, 2\pi] \to \mathbb{R}$ of f. In fact, the difference between $f^*(2\pi)$ and $f^*(0)$ is related to the degree of f; specifically, $f^*(2\pi) - f^*(0) = 2\pi \deg(f)$, as we will show in Theorem 9.31.

Now we are prepared to prove the existence claim of Theorem 9.5. Let $c_n : [0, 2\pi] \to S^1$ be the circle function defined by $c_n(\theta) = e^{in\theta}$.

THEOREM 9.28. *Assume that $f : [0, 2\pi] \to S^1$ is a circle function. There exists $n \in \mathbb{Z}$ such that f is circle homotopic to c_n.*

Proof. Let $f^* : [0, 2\pi] \to \mathbb{R}$ be a lifting of f. Since $f(0) = f(2\pi)$, it follows that $p(f^*(0)) = p(f^*(2\pi))$, and therefore there exists $n \in \mathbb{Z}$ such that $f^*(2\pi) - f^*(0) = 2\pi n$.

Consider the straight-line homotopy $H : [0, 2\pi] \times [0, 1] \to \mathbb{R}$, given by $H(\theta, t) = (1 - t)f^*(\theta) + tn\theta$. The function H is a homotopy between f^* and the function $c_n^* : [0, 2\pi] \to \mathbb{R}$, defined by $c_n^*(\theta) = n\theta$. Define $G : [0, 2\pi] \times [0, 1] \to S^1$ by $G(x, t) = p(H(x, t))$. We claim that G is a circle homotopy between f and c_n.

To begin, we need to show that $G(0, t) = G(2\pi, t)$ for all $t \in [0, 1]$. Thus let $t \in [0, 1]$ be fixed. Note that $H(0, t) = (1 - t)f^*(0)$ and $H(2\pi, t) = (1 - t)f^*(2\pi) + tn2\pi = (1 - t)(2\pi n + f^*(0)) + t2\pi n = (1 - t)f^*(0) + 2\pi n = H(0, t) + 2\pi n$.

Thus $H(0, t)$ and $H(2\pi, t)$ differ by $2\pi n$, and it follows that $p(H(0, t)) = p(H(2\pi, t))$. Therefore $G(0, t) = G(2\pi, t)$, implying that G is a circle homotopy.

Now, $G(\theta, 0) = p(f^*(\theta)) = f(\theta)$ and $G(\theta, 1) = p(n\theta) = e^{in\theta}$. Thus G is a circle homotopy between f and c_n.   ∎

The last step in establishing Theorem 9.5 is to prove the uniqueness result. That is, we need to show that if f is circle homotopic to c_n and c_m, then $n = m$. It suffices to prove that if c_n is circle homotopic to c_m, then $n = m$. To accomplish this, we need another lifting result, this one for homotopies.

THEOREM 9.29. *Let $H : [0, 2\pi] \times [0, 1] \to S^1$ be continuous. Then there exists a function $H^* : [0, 2\pi] \times [0, 1] \to \mathbb{R}$ that is a lifting of H.*

Proof. The approach here is similar to the one used in proof of Theorem 9.27. Again, let $U = S^1 - \{1\}$ and $V = S^1 - \{-1\}$. Then $\{H^{-1}(U), H^{-1}(V)\}$ is an open cover of $[0, 2\pi] \times [0, 1]$. As a consequence of f, the Lebesgue Number Lemma, we can subdivide $[0, 2\pi] \times [0, 1]$ with

$$0 = \theta_0 < \theta_1 < \ldots < \theta_n = 2\pi \text{ and}$$
$$0 = t_0 < t_1 < \ldots < t_m = 1$$

such that each rectangle $[\theta_{i-1}, \theta_i] \times [t_{j-1}, t_j]$ is contained in either $H^{-1}(U)$ or $H^{-1}(V)$. By Theorem 9.24, we can obtain liftings of H restricted to each of these rectangles. The task is to properly choose these liftings and paste them together to obtain the desired lifting H^*.

To simplify the notation, let $R_{i,j}$ be the rectangle $[\theta_{i-1}, \theta_i] \times [t_{j-1}, t_j]$. Begin by defining $G_{1,1} : R_{1,1} \to \mathbb{R}$ to be a lifting of $H|_{R_{1,1}}$. We also denote $G_{1,1}$ by $H^*_{1,1}$. Note that $G_{1,1}(\theta_1, t_0) \in p^{-1}(H(\theta_1, t_0))$. Next, let $G_{2,1} : R_{2,1} \to \mathbb{R}$ be a lifting of $H|_{R_{2,1}}$ satisfying $G_{2,1}(\theta_1, t_0) = G_{1,1}(\theta_1, t_0)$. At this point we have liftings of H to each of the rectangles $R_{1,1}$ and $R_{2,1}$ illustrated in Figure 9.31. These liftings agree at the point (θ_1, t_0), but we need to have them agree over the entire intersection, $R_{1,1} \cap R_{2,1}$, in order to be able to paste them together to have the lifting defined over $R_{1,1} \cup R_{2,1}$.

We claim that, in fact, they do agree over $R_{1,1} \cap R_{2,1}$. The intersection is the set $S = \{\theta_1\} \times [t_0, t_1]$. The restrictions $G_{1,1}|_S$ and $G_{2,1}|_S$ are both liftings of $H|_S$, and they agree at the point (θ_1, t_0). Theorem 9.26 implies that $G_{1,1}|_S = G_{2,1}|_S$, and therefore $G_{1,1}$ and $G_{2,1}$ agree over $R_{1,1} \cap R_{2,1}$. Thus we paste $G_{1,1}$ and $G_{2,1}$ together to obtain a lifting $H^*_{2,1}$ of $H|_{R_{1,1} \cup R_{2,1}}$.
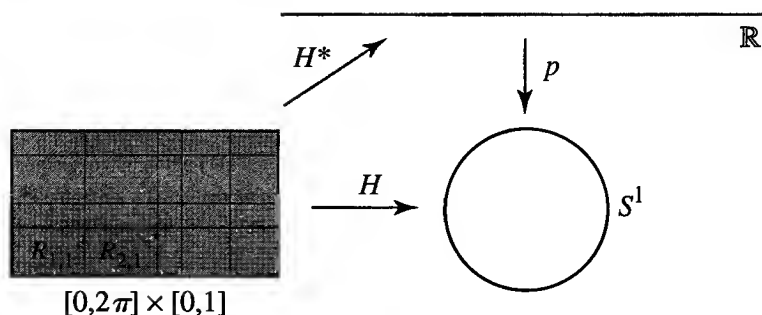
FIGURE 9.31: The homotopy H is lifted one rectangle at a time.

We continue this construction, building H^* across the bottom row of rectangles in the partition of $[0, 2\pi] \times [0, 1]$, then across the next row, and so on. The function $H^*_{n,m}$ is the desired lifting of H. ∎

Next we have the uniqueness result establishing that degree is well defined.

THEOREM 9.30. *If c_n is circle homotopic to c_m, then $n = m$.*

Proof. Let $G : [0, 2\pi] \times [0, 1] \to S^1$ be a circle homotopy with $G(\theta, 0) = c_n(\theta)$ and $G(\theta, 1) = c_m(\theta)$. By Theorem 9.29 there exists a function $G^* : [0, 2\pi] \times [0, 1] \to \mathbb{R}$ that is a lifting of G.

Since $G(\theta, 0) = c_n(\theta)$, it follows that $G^*|_{[0,2\pi]\times\{0\}}$ is a lifting of c_n. The function $c_n^* : [0, 2\pi] \to \mathbb{R}$, defined by $c_n^*(\theta) = n\theta$, is also a lifting of c_n. Theorem 9.26 implies that there exists $d \in \mathbb{Z}$ such that $G^*(\theta, 0) = n\theta + 2\pi d$ for all $\theta \in [0, 2\pi]$. Therefore

$$G^*(2\pi, 0) - G^*(0, 0) = 2\pi n + 2\pi d - 2\pi d = 2\pi n. \qquad (9.4)$$

Similarly, we obtain

$$G^*(2\pi, 1) - G^*(0, 1) = 2\pi m. \qquad (9.5)$$

Now, G being a circle homotopy implies that the functions $G|_{\{0\}\times[0,1]}$ and $G|_{\{2\pi\}\times[0,1]}$, considered as functions mapping $[0, 1]$ to S^1, are identical. Therefore $G^*|_{\{0\}\times[0,1]}$ and $G^*|_{\{2\pi\}\times[0,1]}$ are liftings of the same function. Theorem 9.25 implies that there exists $b \in \mathbb{Z}$ such that $G^*(2\pi, t) = G^*(0, t) + 2\pi b$ for all $t \in [0, 1]$. It follows that $G^*(2\pi, 1) - G^*(0, 1) = G^*(2\pi, 0) - G^*(0, 0)$, which, by Equations 9.4 and 9.5, implies that $2\pi m = 2\pi n$, and therefore $m = n$. ∎

Through Theorems 9.28 and 9.30 we have proven Theorem 9.5, establishing that degree is a well-defined property of circle functions. By a proof similar to the proof of Theorem 9.30, we can establish the following theorem relating degree to the endpoints of liftings of circle functions:

THEOREM 9.31. *Let* $f : [0, 2\pi] \to S^1$ *be a circle function, and let* $f^* : [0, 2\pi] \to \mathbb{R}$ *be a lifting of* f. *Then* $f^*(2\pi) - f^*(0) = 2\pi \deg(f)$.

Proof. See Exercise 9.26.   ∎

Exercises for Section 9.6

9.24. Show that if we drop the hypothesis that X is connected in Theorem 9.25, then the result need not hold. That is, find a continuous function $f : X \to S^1$, with disconnected domain X, such that f has liftings $g, h : X \to \mathbb{R}$ that do not differ by a constant.

9.25. **Prove Theorem 9.26:** Let $f : X \to S^1$ be continuous, X be connected, and $g, h : X \to \mathbb{R}$ be liftings of f. If there exists $x_0 \in X$ such that $g(x_0) = h(x_0)$, then $g(x) = h(x)$ for all $x \in X$.

9.26. **Prove Theorem 9.31:** Let $f : [0, 2\pi] \to S^1$ be a circle function, and let $f^* : [0, 2\pi] \to \mathbb{R}$ be a lifting of f. Then $f^*(2\pi) - f^*(0) = 2\pi \deg(f)$. (Hint: Let G be a circle homotopy between f and c_n where $n = \deg(f)$, $G(\theta, 0) = f(\theta)$, and $G(\theta, 1) = c_n(\theta)$. Let G^* be a lifting of G. Argue as in the proof of Theorem 9.30 and show that $G^*(2\pi, 0) - G^*(0, 0) = 2\pi n$. Then use Theorem 9.25 to obtain the desired result.)

9.27. Prove that if $f : S^1 \to S^1$ is a homeomorphism, then $\deg(f) = \pm 1$.

Supplementary Exercises: The Borsuk–Ulam Theorem and the Ham Sandwich Theorem

We introduced the Borsuk–Ulam Theorem in Section 6.3, where we proved that if $f : S^2 \to \mathbb{R}$ is continuous, then there exists $x \in S^2$ such that $f(x) = f(-x)$ (viewing $x \in S^2$ as a unit vector based at the origin in $\mathbb{R}^3$). We indicated then that we could do better, specifically that the same result holds with range $\mathbb{R}^2$ rather than $\mathbb{R}$. As mentioned in Section 6.3, this implies that there must be a point on the surface of the Earth that has exactly the same temperature and barometric pressure as its antipodal point. We are now in a position to prove this stronger result; we do so through the following exercises.

A circle function $f : S^1 \to S^1$ is said to be antipode-preserving if f maps antipodal points to antipodal points. That is, if we view f as complex-valued, then for all complex z in S^1 we have $f(-z) = -f(z)$. (See Figure 9.32.) We can also express
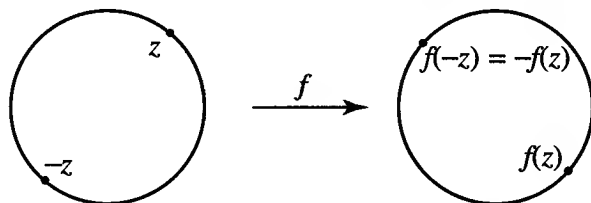
FIGURE 9.32: Antipode-preserving functions map antipodal points to antipodal points.

this relationship as $f(e^{i(\theta+\pi)}) = e^{i\pi} f(e^{i\theta})$. When the circle function is represented as a map $f : [0, 2\pi] \rightarrow S^1$ such that $f(0) = f(2\pi)$, then to be antipode-preserving f must satisfy $f(\theta + \pi) = -f(\theta) = e^{i\pi} f(\theta)$ for all $\theta \in [0, \pi]$.

This requirement yields $f(2\pi) = -f(\pi) = -(-f(0)) = f(0)$. Of course, $f(2\pi) = f(0)$ is required of circle functions by their definition, and therefore the requirements for being a circle function and for being antipode-preserving are consistent with each other.

Assume that $f : [0, 2\pi] \rightarrow S^1$ is an antipode-preserving circle function, and let $f^* : [0, 2\pi] \rightarrow \mathbb{R}$ be a lifting of f.

SE 9.28. Prove that there exists $n \in \mathbb{Z}$ such that $f^*(\theta + \pi) = f^*(\theta) + (2n + 1)\pi$ for all $\theta \in [0, \pi]$.

SE 9.29. Prove that $f^*(2\pi) - f^*(0) \neq 0$, and conclude that antipode-preserving circle functions have nonzero degree.

Now we want to expand the concept of antipode-preserving to include functions $f : S^n \rightarrow S^m$. In order to do so most conveniently, we represent points in S^n using unit vectors x based at the origin in $\mathbb{R}^{n+1}$. In this case, the requirement for a function $f : S^n \rightarrow S^m$ to be antipode-preserving is that $f(-x) = -f(x)$ for all $x \in S^n$.

THEOREM 9.32. *There is no continuous antipode-preserving function $f : S^2 \rightarrow S^1$.*

SE 9.30. Prove Theorem 9.32. (Hints: Assume that there is an antipode-preserving function $f : S^2 \rightarrow S^1$ and let f_0 be the circle function obtained by restricting f to the equator. Show that you can apply Exercise SE 9.29 to f_0. Then restrict f to the upper half of S^2, including the equator, and argue that you can apply Theorem 9.9 to f_0. You should be able to derive a contradiction resulting from these two views of f_0.)

THEOREM 9.33. *The Borsuk–Ulam Theorem.* *Let $f : S^2 \rightarrow \mathbb{R}^2$ be continuous. There exists $x \in S^2$ such that $f(x) = f(-x)$.*

SE 9.31. Prove Theorem 9.33. (Hint: Assume no such x exists and consider the function defined by $h(x) = \frac{f(x)-f(-x)}{|f(x)-f(-x)|}$.)

Along with the simple meteorological application already mentioned, another interesting consequence of the Borsuk–Ulam Theorem is the Ham Sandwich Theorem. This states that if you have three solids in $\mathbb{R}^3$, then there is a plane in $\mathbb{R}^3$ that simultaneously cuts each solid in half. (See Figure 9.33.) Thus if you make a sandwich from a piece of bread, a slice of ham, and another piece of bread, then no matter how you arrange the bread and ham, you can cut through the sandwich with a single, straight knife cut and divide each piece of bread and the ham exactly in half. The Ham Sandwich Theorem plays a role in fair-division problems in applications of mathematics in the social sciences. These problems address, for example, the fair distribution of resources among multiple parties vying for them. (See [Hil], for example.)

FIGURE 9.33: There is a plane that cuts each of the three solids in half.

We now discuss an approach to proving the Ham Sandwich Theorem. Let σ_1, σ_2, and σ_3 be three solids in $\mathbb{R}^3$. By "solid" we mean a bounded set with a well-defined volume, a notion from measure theory whose details we leave unspecified. We define a function $d : S^2 \to \mathbb{R}^2$ that is used to prove the existence of a plane bisecting each of the solids. Each point x in S^2 is treated as a unit vector in $\mathbb{R}^3$, describing a direction in $\mathbb{R}^3$. Fix $x \in S^2$. Consider all of the planes that have x as a normal vector.

SE 9.32. Discuss why there must be a plane $P_1(x)$, with normal vector x, that cuts solid σ_1 in half. (Hint: For each plane P with normal vector x, consider the quantity V_P that indicates the percent of the volume of σ_1 that lies on the side of P in the direction of the normal vector x.)

Similarly, there must be planes $P_2(x)$ and $P_3(x)$ with normal vector x that cut solids σ_2 and σ_3 in half, respectively. We show that there must be an $x^* \in S^2$ such that $P_1(x^*) = P_2(x^*) = P_3(x^*)$, and therefore this common plane simultaneously cuts each solid in half.

Let $d_1(x)$ be the distance from plane $P_3(x)$ to plane $P_1(x)$ if plane $P_1(x)$ lies on the side of $P_3(x)$ in the direction of the normal vector x, and let $d_1(x)$ be the negative of the distance from plane $P_3(x)$ to plane $P_1(x)$ if plane $P_1(x)$ lies on the side of $P_3(x)$ opposite the direction of the normal vector x. Define $d_2(x)$ similarly in terms of the distance from $P_3(x)$ to $P_2(x)$. We assume that the function $d : S^2 \to \mathbb{R}^2$ defined by $d(x) = (d_1(x), d_2(x))$ is continuous.

SE 9.33. Prove that $d(-x) = -d(x)$ for each $x \in S^2$.

SE 9.34. Use the Borsuk–Ulam Theorem, in conjunction with Exercise SE 9.33, to prove that there exists $x^* \in S^2$ such that $d(x^*) = (0, 0)$. Show that the Ham Sandwich Theorem follows.

Fixed Point Theorems and Applications

Let $f : X \to X$ be a function mapping a topological space X to itself. If $f(x) = x$ for some x in X, then x maps to itself under f and is known as a fixed point of f. Fixed points and theorems about them play an important role in many areas of pure and applied mathematics. Within topology itself there is a large body of work on fixed point results. One of the most well known is the Brouwer Fixed Point Theorem. It states that every continuous function from an n-ball B^n to itself must have a fixed point. We proved the One-Dimensional Brouwer Fixed Point Theorem as a consequence of the Intermediate Value Theorem in Section 6.3. Here, in Section 10.1, we prove the Two-Dimensional Brouwer Fixed Point Theorem as a consequence of the Two-Dimensional No Retraction Theorem. In Section 10.2, we present an application of the Brouwer Fixed Point Theorem to prove the existence of equilibrium price distributions in a pure exchange economy. In Section 10.3, we present a generalization of the Brouwer Fixed Point Theorem, known as Kakutani's Fixed Point Theorem, which applies to set-valued functions. Finally, in Section 10.4, we apply Kakutani's Fixed Point Theorem to prove the existence of Nash equilibria in game theory. This is one of the most important results in the field of game theory, as it demonstrates that there is a choice of strategies that optimizes the expected outcome for all players of a game.

10.1 The Brouwer Fixed Point Theorem

Imagine taking two pieces of the same-sized paper and laying one piece on top of the other. Every point on the top sheet of paper is associated with some point right below it on the other sheet. Now take the top sheet of paper and crumple it up into a ball without ripping it. Place the crumpled ball back on top of the bottom sheet of paper. Somewhere on the crumpled ball of paper there is a point that is sitting directly above the same point on the bottom sheet of paper that it sat above before the crumpling took place. (See Figure 10.1.) This is an application of the Two-Dimensional Brouwer Fixed Point Theorem, which we prove in this section.

DEFINITION 10.1. *Let $f : X \to X$. A point $x \in X$ is said to be a **fixed point** of f if $f(x) = x$. A topological space X is said to have the **fixed point property** if every continuous function $f : X \to X$ has a fixed point.*

EXAMPLE 10.1. The space $[-1, 1]$ has the fixed point property by the One-Dimensional Brouwer Fixed Point Theorem.
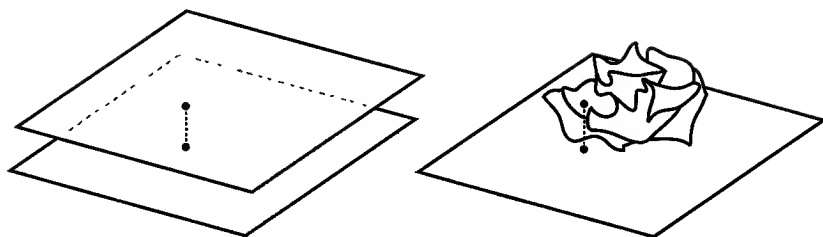
FIGURE 10.1: After the paper crumpling, some point still sits above the same point it sat above prior to crumpling.

EXAMPLE 10.2. The real line $\mathbb{R}$ does not have the fixed point property since there exist continuous functions from $\mathbb{R}$ to $\mathbb{R}$ that do not fix any point. The function $f : \mathbb{R} \to \mathbb{R}$ given by $f(x) = x + 1$ is an example.

EXAMPLE 10.3. The circle S^1 does not have the fixed point property since the continuous function $f : S^1 \to S^1$, given by rotating each point on the circle through an angle of π, does not have any fixed points.

The fixed point property is a topological property, meaning that if spaces X and Y are homeomorphic, then X has the fixed point property if and only if Y does. (See Exercise 10.3.) Therefore all closed bounded intervals $[a, b]$ with the standard topology have the fixed point property since $[-1, 1]$ does. Also, all open intervals $(-\infty, b)$, (a, b), and (a, ∞) with the standard topology do not have the fixed point property since $\mathbb{R}$ does not.

One of the most useful theorems in topology is the Brouwer Fixed Point Theorem. It says that for each n, the n-ball B^n has the fixed point property. This theorem is named after the Dutch mathematician L. E. J. Brouwer (1881–1966) who proved the general n-dimensional version of the theorem in 1912. We show here that the Two-Dimensional Brouwer Fixed Point Theorem is equivalent to the Two-Dimensional No Retraction Theorem (with a proof that actually carries over to establish this equivalence for all n). Since we established the Two-Dimensional No Retraction Theorem in Section 9.2, the Two-Dimensional Brouwer Fixed Point Theorem follows.

In a set of supplementary exercises in this section, we present a second proof of the Two-Dimensional Brouwer Fixed Point Theorem, one that involves the degree of circle functions, but does not rely on the No Retraction Theorem.

THEOREM 10.2. *The Two-Dimensional Brouwer Fixed Point Theorem.* *Every continuous function $f : D \to D$, mapping the disk to itself, has a fixed point.*

As already indicated, our approach to proving this theorem is to show its equivalence to the Two-Dimensional No Retraction Theorem. That is done via the following theorem:

THEOREM 10.3. *The disk D, as a subspace of* $\mathbb{R}^2$*, has the fixed point property if and only if there is no retraction from D onto its boundary* S^1.

Proof. To begin, assume that there is a retraction $r : D \rightarrow S^1$. Consider the map $q : S^1 \rightarrow D$, defined by $q(x) = -x$, where we consider x as a vector in the plane. The function $q \circ r : D \rightarrow D$ is continuous and has no fixed point. Therefore, if there is a retraction $r : D \rightarrow S^1$, then D does not have the fixed point property.

On the other hand, assume that a continuous function $f : D \rightarrow D$ has no fixed point. We show that there is a retraction $r : D \rightarrow S^1$. Define $r : D \rightarrow S^1$ as follows. First, take the ray in $\mathbb{R}$ running from $f(x)$ through x. Such a ray is well defined since f has no fixed point. Let $r(x)$ be the point where the ray intersects S^1, as illustrated in Figure 10.2. Clearly, r maps D to S^1 and $r(x) = x$ for all $x \in S^1$. It will follow that r is a retraction once the continuity of r is established.
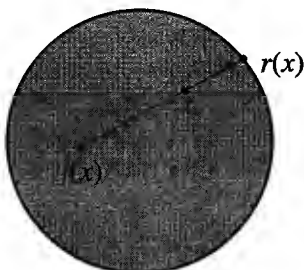
FIGURE 10.2: Defining r from the disk to the bounding circle.

To prove that r is continuous, let U be open in S^1 and x be a point in $r^{-1}(U)$. We show that there is an open set V containing x such that $r(V) \subset U$, and therefore $r^{-1}(U)$ is open. To begin, we choose small open balls O_1 and O_2 centered at $f(x)$ and x, respectively, such that every ray beginning in O_1 and passing through O_2 intersects S^1 in the set U. (See Figure 10.3.) Since f is continuous, we can find an open set V, containing x and contained in O_2, such that $f(V) \subset O_1$. Thus, for all $v \in V$, the ray beginning at $f(v)$ and passing through v intersects S^1
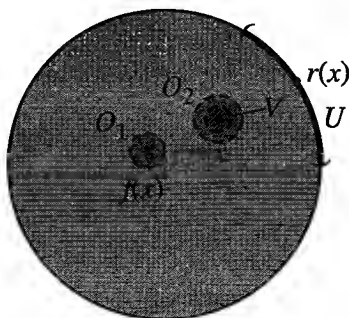
FIGURE 10.3: There exists an open set V containing x such that $f(V) \subset U$.

in U. Therefore $r(v) \in U$, and it follows that r is continuous. Hence we have established that if D does not have the fixed point property, then there is a retraction $r : D \to S^1$, and the proof of the theorem is complete. ∎

The Two-Dimensional No Retraction Theorem and Theorem 10.3 immediately imply Theorem 10.2, and therefore we have established the Two-Dimensional Brouwer Fixed Point Theorem.

The approach to the proof of Theorem 10.3 carries through if we replace the disk D and the circle S^1 with the n-ball B^n and the $(n-1)$-sphere S^{n-1}, respectively, and in this way we can show that the n-Dimensional Brouwer Fixed Point Theorem is equivalent to the n-Dimensional No Retraction Theorem.

EXAMPLE 10.4. Suppose we take a map of New England and place it on the ground anywhere within New England, as in Figure 10.4. We assume that New England is topologically equivalent to a disk, and we refer to it as N. Let f be the function assigning to each point in New England the point on the map corresponding to it. We can view f as a continuous function from N to itself. Therefore f must have a fixed point, from which it follows that there must be a point on the map that corresponds exactly to the point on the ground directly beneath it. If, as a guide to travelers, we leave the map where it is, then on the map we would indicate the location of the fixed point with an arrow labeled "You Are Here!"

FIGURE 10.4: The function from New England to the map must have a fixed point.

Exercises for Section 10.1

10.1. Show that each of the following spaces does not have the fixed point property:
  (a) The interval $(0, 1]$
  (b) The torus
  (c) The figure-eight space obtained by taking two circles and gluing them together at a point on each
  (d) The sphere

10.2. Prove that if a topological space X has the fixed point property, then X is connected.

10.3. Show that if X has the fixed point property and Y is homeomorphic to X, then Y has the fixed point property.

10.4. Show that if X has the fixed point property and A is a retract of X, then A has the fixed point property.

10.5. Show that if X does not have the fixed point property, then for all Y, the product space $X \times Y$ does not have the fixed point property.

10.6. Prove the One-Dimensional Brouwer Fixed Point Theorem from the Two-Dimensional Brouwer Fixed Point Theorem.

10.7. Consider the topological graphs of the letters A–E as illustrated in Figure 10.5. Determine which of these spaces has the fixed point property. (Hint: For the letter E, you may want to use Exercise 10.4.)
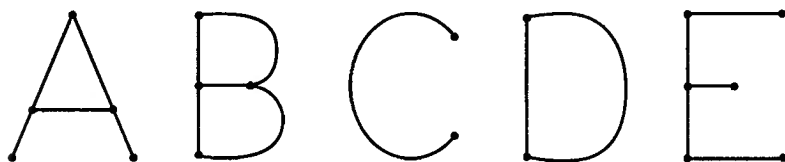
FIGURE 10.5: Which of these letters has the fixed point property?

Supplementary Exercises: Another Approach to the Two-Dimensional Brouwer Fixed Point Theorem

In these exercises, we develop another proof of the Two-Dimensional Brouwer Fixed Point Theorem. In this case the proof is based directly on the degree of circle functions and does not rely on the No Retraction Theorem. In what follows, we assume that points x in S^1, D, and so on, are vectors based at the origin in the plane. We begin with a lemma.

LEMMA 10.4. *Let $g : S^1 \to \mathbb{R}^2 - \{O\}$ be continuous, and assume that the function $h : S^1 \to S^1$, defined by $h(x) = \frac{g(x)}{|g(x)|}$, has degree 1. Then there exists $x \in S^1$ such that $g(x) = ax$ for some $a > 0$.*

The lemma indicates that some $x \in S^1$ (again, considered as a vector based at the origin) maps to a positive scalar multiple of itself under g.

> **Proof.** Define $G : S^1 \times [0, 1] \to \mathbb{R}^2$ by $G(x, t) = (1 - t)g(x) - tx$. The function G defines a straight-line homotopy from g (considered to have range $\mathbb{R}^2$) to the function $a(x) = -x$, which maps each $x \in S^1$ to its antipodal point $-x \in S^1$.

SE 10.8. Show that there exists $(x, t) \in S^1 \times (0, 1)$ such that $G(x, t) = 0$. (Hint: Assume that no such point (x, t) exists, and consider the homotopy of circle functions $H(x, t) : S^1 \times [0, 1] \to S^1$, defined by $H(x, t) = \frac{G(x,t)}{|G(x,t)|}$.)

SE 10.9. Show that if there exists $(x, t) \in S^1 \times (0, 1)$ such that $G(x, t) = 0$, then $g(x) = ax$ for some $a > 0$, thereby proving the lemma. ∎

Circle functions do not necessarily have fixed points. A simple rotation of a circle by π leaves no point fixed. This rotation function has degree 1. Interestingly, that is the only degree for which a circle function can have no fixed point. For circle functions whose degree is not equal to 1, we have the following theorem:

THEOREM 10.5. *Let $f : S^1 \to S^1$ be a circle function with $\deg(f) \neq 1$. Then there exist $x', x'' \in S^1$ such that $f(x') = x'$ and $f(x'') = -x''$.*

Thus if the degree of a circle function is not equal to 1, then the function has a fixed point and a point that maps to its antipodal point.

SE 10.10. Prove Theorem 10.5. (Hint: Regard $f : S^1 \to S^1$ as a function mapping into $\mathbb{R}^2 - \{O\}$, and use Lemma 10.4 to prove the existence of the fixed point. Then consider the function $h : S^1 \to S^1$ defined by $h(x) = f(-x)$ and apply the first assertion of the theorem to h.)

Now, we proceed with the Two-Dimensional Brouwer Fixed Point Theorem. Let D be the disk in the plane.

THEOREM 10.6. *Let $f : D \to D$ be continuous. Then there exists $x \in D$ such that $f(x) = x$.*

> **Proof.** Assume that this is not the case. Let $k : D \to \mathbb{R}^2 - \{O\}$ be the function defined by $k(x) = f(x) - x$. The function k maps to $\mathbb{R}^2 - \{O\}$ since we are assuming that f has no fixed point.

SE 10.11. Indicate why Lemma 10.4 applies to the function $k|_{S^1} : S^1 \to \mathbb{R}^2 - \{O\}$. (Hint: Show that the degree of the circle function involved is 0.)

Therefore there exists $x^* \in S^1$ such that $k(x^*) = ax^*$ for some $a > 0$.

SE 10.12. Show that $f(x^*) \in D$, thereby arriving at a contradiction and completing the proof of the theorem. ∎

10.2   *An Application to Economics*

In this section, we show how fixed point theory can demonstrate the existence of economic equilibria. Suppose we have an economy consisting of a finite number of individuals who have certain items to buy and sell. We assume that there is neither consumption of the items nor new production of them. Therefore individuals who wish to have more of a particular item must sell some of their current stock of other items in order to purchase the item they desire. This is called a pure exchange economy.

The price of the items is determined by supply and demand, what in 1776 philosopher and economist Adam Smith (1723–1790) called the "invisible hand." If there is more demand for a product than there is supply, the product's price goes up. If there is excess supply and little demand, the product's price goes down. We would like to know whether there is a choice of prices that balances supply and demand, an equilibrium point where prices stabilize and where all consumers are happy with their particular bundle of goods.

In 1932, John Von Neumann (1903–1957) gave a seminar at Princeton entitled "On a System of Economic Equations and a Generalization of the Brouwer Fixed Point Theorem." In it, he outlined how fixed point theory could be utilized to prove the existence of equilibria in economic models. Generalizations and applications of this concept have resulted in Nobel Prizes in economics for Kenneth Arrow in 1972 and Gerard Debreu (1921–2004) in 1983. Applications to game theory, as discussed in Section 10.4, also led to a Nobel Prize in economics for the mathematician John Nash in 1994.

Let us look at a particularly simple example as a starting point. Suppose we have an economy with only three items available. These are cashmere, butter, and gunpowder. The total quantity of each, in pounds, is called the supply, and the three supplies are denoted S_C, S_B, and S_G, respectively. We keep track of the supply of each item in a vector called the supply vector $\mathbf{S} = (S_C, S_B, S_G)$.

There are n individuals trading in this economy. We denote them by the integers $1, 2, \ldots, n$. Each individual starts with some number of each item, collectively called his or her bundle of goods. Individual i has a bundle $b = (b_C^i, b_B^i, b_G^i)$, a three-dimensional vector giving the quantity of each item in the corresponding component. At any given time, the sum of all of the bundle vectors over the individuals in the economy gives the supply vector,

$$\mathbf{S} = \sum_i \mathbf{b}^i.$$

We are assuming that none of the items are consumed or destroyed over time, so $\mathbf{S}$ is constant. Everyone is simply stockpiling their goods and desirous of a particular mix, depending on the going prices. If an individual wants more butter and less cashmere, she can sell some of her cashmere to someone else at the going price and purchase additional butter. However, if everyone is trying to get rid of cashmere, the price of cashmere is driven down.

We denote the going prices per pound for cashmere, butter, and gunpowder by p_C, p_B, and p_G respectively. We then represent this set of prices by a single price vector $\mathbf{p} = (p_C, p_B, p_G)$.

In fact, it is only the relative price of these items that matters. The particular cost of any item is irrelevant, since it is the ratio of costs that determines, for instance, how much butter can be purchased when a pound of cashmere is sold.

Therefore we divide each of the individual prices by the sum of the prices $p_C + p_B + p_G$ to "normalize" the situation. We denote the resultant prices by

$$p_C' = \frac{p_C}{p_C + p_B + p_G},$$

$$p_B' = \frac{p_B}{p_C + p_B + p_G}, \text{ and}$$

$$p_G' = \frac{p_G}{p_C + p_B + p_G}.$$

These normalized prices have the advantage of summing to 1. To simplify notation, we drop the primes on the price vectors, but keep in mind our assumption that the prices have been normalized. We then keep track of these

vectors in our overall price vector $\mathbf{p} = (p_C, p_B, p_G)$ such that

$$p_C + p_B + p_G = 1.$$

In addition, we assume that none of these items is so despised by a consumer that he might pay you to take it away. In other words, the prices are never negative, so

$$p_C \geq 0, \quad p_B \geq 0, \text{ and } p_G \geq 0.$$

When the price of an item is 0, there is no demand for that item. Each consumer would rather have the other two items, no matter how expensive they are.

We can graph the set of possible price vectors in 3-space. The equation $p_C + p_B + p_G = 1$ defines a plane. That $p_C \geq 0$, $p_B \geq 0$, and $p_G \geq 0$ limits us to that part of the plane that lies in the first octant of 3-space, yielding a triangle T, as shown in Figure 10.6. Each point in the triangle corresponds to a triple of prices, one for each of our three goods. It will be important for us, in applying the Brouwer Fixed Point Theorem, that T is topologically equivalent to a disk.
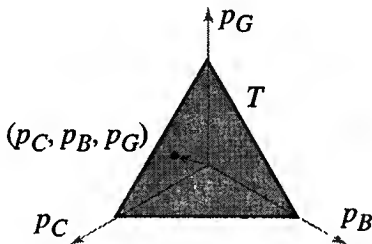
FIGURE 10.6: The triangle T of possible price vectors (p_C, p_B, p_G) for our three goods.

Given a particular price vector $\mathbf{p}$, each individual in the economy has a certain amount of "wealth" in the form of the value of his or her bundle of goods. For the ith consumer, this wealth is given by

$$w^i(\mathbf{p}) = \mathbf{p} \cdot \mathbf{b}^i.$$

An individual's wealth can change depending on the current price vector, so we represent it as a function of $\mathbf{p}$.

At a particular price vector $\mathbf{p}$, individual i's current bundle of goods may not be her optimal choice. She might want to trade certain goods for others. We keep track of her preferences for an optimal bundle at price $\mathbf{p}$ with a demand vector $\mathbf{d}^i(\mathbf{p})$, which gives the combination of items she would like to have at these prices. We assume she spends all of her wealth in exchanging her bundle $\mathbf{b}^i$ to obtain her desired bundle $\mathbf{d}^i(\mathbf{p})$. We express this with the following equation:

$$w^i(\mathbf{p}) = \mathbf{p} \cdot \mathbf{b}^i = \mathbf{p} \cdot \mathbf{d}^i(\mathbf{p}). \tag{10.1}$$

EXAMPLE 10.5. Suppose Wilma has a pound of cashmere, a pound of butter, and three pounds of gunpowder. So she has a supply bundle $\mathbf{b}^W = (1, 1, 3)$. At the price vector $\mathbf{p} = (\frac{1}{6}, \frac{3}{6}, \frac{2}{6})$, she has a net worth of $\mathbf{p} \cdot \mathbf{b}^W = 5/3$. Furthermore, assume that at this price vector, she would prefer a goods distribution as given by the demand vector $\mathbf{d}^W(\mathbf{p}) = (2, 2, 1)$. As required by Equation 10.1, this demand vector is constrained so that the total cost of her bundle at these prices, $\mathbf{p} \cdot \mathbf{d}^W(\mathbf{p})$, is equal to her net worth 5/3.

On the other hand, suppose Dexter has bundle $\mathbf{b}^D = (1, 2, 4)$, giving him a net worth of 5/2. For the same price vector $\mathbf{p}$, he has a demand vector of $\mathbf{d}^D(\mathbf{p}) = (6, 2, \frac{3}{2})$. Notice that Dexter seems to like cashmere more than Wilma since he wishes to convert proportionately more of his wealth to cashmere than she does.

Now, if the prices change, the demand vectors can change as well. For example, if the price vector is $\mathbf{p}' = (\frac{4}{10}, \frac{1}{10}, \frac{5}{10})$, then Wilma's net worth increases to $\mathbf{p}' \cdot \mathbf{b}^W = 2$. Seeing such a low price for butter, she may wish to stock up on it and choose a demand vector $\mathbf{d}^W(\mathbf{p}') = (1, 11, 1)$. Here too, her net worth does not change (it remains at 2), but she has converted much of it into a large supply of low-priced butter.

Each consumer i has a vector-valued demand function $\mathbf{d}^i$ which maps price vectors in T to the consumer's corresponding demand vector. Adding together the individual demand vectors for all of the consumers yields the demand vector $\mathbf{D}(\mathbf{p})$ for the entire economy at the given price vector $\mathbf{p}$. Thus,

$$\mathbf{D}(\mathbf{p}) = \sum_i \mathbf{d}^i(\mathbf{p}).$$

We assume that this vector-valued function $\mathbf{D}$ is continuous in the sense that small changes in the price vector cause small changes in overall demand.

Since $\mathbf{p} \cdot \mathbf{d}^i(\mathbf{p})$ is the net worth of individual i given price vector $\mathbf{p}$, the dot product $\mathbf{p} \cdot \mathbf{D}(\mathbf{p})$ equals the total wealth of the community given those prices. Note that

$$\mathbf{p} \cdot \mathbf{D}(\mathbf{p}) = \sum_i \mathbf{p} \cdot \mathbf{d}^i(\mathbf{p}) = \sum_i \mathbf{p} \cdot \mathbf{b}^i = \mathbf{p} \cdot \mathbf{S}.$$

This yields

Walras's Law: $\mathbf{p} \cdot \mathbf{D}(\mathbf{p}) = \mathbf{p} \cdot \mathbf{S}.$

This relationship is named for Leon Walras (1829–1910), one of the first economists to put the field on a mathematical footing.

If the coordinate of $\mathbf{D}(\mathbf{p})$ corresponding to cashmere is greater than the coordinate of the supply vector $\mathbf{S}$ corresponding to cashmere, then at this price there is more demand for cashmere than there is supply. This will drive up the price of cashmere.

However, if at the current prices, everyone can obtain their optimal choice of goods and would not trade any of their current quantity of each item, then the economy is in equilibrium.

Let v_j denote the jth component of a vector $\mathbf{v}$.

DEFINITION 10.7. *A price vector* $\mathbf{p}$ *is an* ***equilibrium price vector*** *if* $D_j(\mathbf{p}) \leq S_j$ *for all* j.

For such a price vector, the supply of every item exceeds the demand. Everyone can obtain their optimal choice of goods. No one desires further exchanges. Although the demand for an item may be strictly less than the total supply of that item, this does not necessarily drive down the price of that item. Since the prices are relative and everyone may have enough of the other products, the price of the item in excess supply need not fall.

Is there an equilibrium price vector? Can the economy be in balance? We will see that the answer is yes using the Brouwer Fixed Point Theorem.

To do so, we define a function $\mathbf{f} : T \to T$ that maps the set of possible price vectors to itself and tells us how the demand for items causes prices to change.

To see how prices change, we consider the following vector:

DEFINITION 10.8. *The excess demand vector is defined by* $\mathbf{E}(\mathbf{p}) = \mathbf{D}(\mathbf{p}) - \mathbf{S}$.

The coordinates of the vector $\mathbf{E}(\mathbf{p})$ tell us whether or not there is more demand or more supply for the items at the prices in the vector $\mathbf{p}$. If a coordinate of $\mathbf{E}(\mathbf{p})$ is positive, we expect the price of the corresponding item to rise since there is more demand than supply. People want more of this product at this price than is available. If the coordinate is negative, however, conditions could drive down the price of this item, as there is more supply than demand at this price.

Using the excess demand vector, Walras's Law can now be rewritten:

$$\textbf{Second form of Walras's Law: } \mathbf{p} \cdot \mathbf{E}(\mathbf{p}) = 0. \qquad (10.2)$$

Given a price vector $\mathbf{p}$, the vector $\mathbf{E}(\mathbf{p})$ points in the direction that we expect the prices to move from $\mathbf{p}$. From it we can construct a function $\mathbf{f}$ that takes each price vector $\mathbf{p}$ in T to another price vector $\mathbf{f}(\mathbf{p})$ in T toward which we would expect the vector $\mathbf{p}$ to move. Hence we will have created a continuous map from T back to itself. The Two-Dimensional Brouwer Fixed Point Theorem tells us that this function must have a fixed point. In other words, there is a price vector $\mathbf{p}$ such that there is no incentive to trade goods when the prices are at $\mathbf{p}$. At that price, the entire economy is in equilibrium.

The rest of this section is devoted to making this idea mathematically precise by appropriately defining $\mathbf{f} : T \to T$ and proving that a fixed point of $\mathbf{f}$ corresponds to an equilibrium price vector.

We begin by defining a function $\mathbf{f}^* : T \to \mathbb{R}^3$ by $\mathbf{f}^*(\mathbf{p}) = \mathbf{p} + \mathbf{E}(\mathbf{p})$. This function gives us an idea of how prices should move due to excess supply or demand.

However, there is a fundamental problem with this definition of $\mathbf{f}^*$. It does not necessarily send price vectors in T back to price vectors in T. The vectors that result might have negative entries, and their coordinates do not necessarily sum to 1. We need to remedy these problems so that we can be in a position to apply the Brouwer Fixed Point Theorem.

In order to make all of the entries nonnegative, we define $\mathbf{v}^+$ to be the vector obtained from a vector $\mathbf{v}$ by changing all negative entries into 0 entries. Then we define our new, improved function to be $\mathbf{f}^{**}(\mathbf{p}) = (\mathbf{p} + \mathbf{E}(\mathbf{p}))^+$. The function $\mathbf{f}^{**}$ maps T to the first octant in $\mathbb{R}^3$, but $\mathbf{f}^{**}(\mathbf{p})$ need not lie in T, since its entries do not necessarily sum to 1. However, by "normalizing" $\mathbf{f}^{**}(\mathbf{p})$, which is to say, dividing the resulting vector by the sum of its entries, we obtain a vector in T. It is important to note that at least one coordinate of $\mathbf{p} + \mathbf{E}(\mathbf{p})$ is positive, and therefore in normalizing $\mathbf{f}^{**}(\mathbf{p})$ we are not dividing by 0. (See Exercise 10.15.)

With $(\mathbf{p} + \mathbf{E}(\mathbf{p}))_j^+$ representing the jth entry of $(\mathbf{p} + \mathbf{E}(\mathbf{p}))^+$, we let $\beta(\mathbf{p}) = \sum_{j=1}^3 (\mathbf{p} + \mathbf{E}(\mathbf{p}))_j^+$, and we define our desired function, mapping T to itself, as follows:

DEFINITION 10.9. *The price change function* $\mathbf{f} : T \to T$ *is defined by*

$$\mathbf{f}(\mathbf{p}) = \frac{(\mathbf{p} + \mathbf{E}(\mathbf{p}))^+}{\beta(\mathbf{p})}.$$

Figure 10.7 depicts an example of the vectors $\mathbf{p}$, $\mathbf{E}(\mathbf{p})$, $\mathbf{p} + \mathbf{E}(\mathbf{p})$, $(\mathbf{p} + \mathbf{E}(\mathbf{p}))^+$, and $\mathbf{f}(\mathbf{p})$ for a two-dimensional situation, as opposed to the three-dimensional one we have been considering. The second form of Walras's Law implies that $\mathbf{p}$ and $\mathbf{E}(\mathbf{p})$ must be perpendicular.
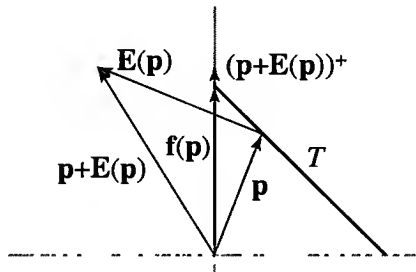
FIGURE 10.7: The vectors $\mathbf{p}$, $\mathbf{E}(\mathbf{p})$, $\mathbf{p} + \mathbf{E}(\mathbf{p})$, $(\mathbf{p} + \mathbf{E}(\mathbf{p}))^+$, and $\mathbf{f}(\mathbf{p})$.

The fact that $\mathbf{D}(\mathbf{p})$ is continuous implies that the price change function $\mathbf{f}(\mathbf{p})$ is also continuous.

Now, we are ready to apply the Two-Dimensional Brouwer Fixed Point Theorem. Since $\mathbf{f} : T \to T$ is a continuous function and T is a disk, there must be a price vector $\mathbf{p}^*$ such that $\mathbf{f}(\mathbf{p}^*) = \mathbf{p}^*$. That is to say, there is some price vector that is fixed by $\mathbf{f}$. We show, in fact, that such a vector is the equilibrium price vector we seek.

THEOREM 10.10. *The fixed points of the price change function* $\mathbf{f}$ *are equilibrium price vectors.*

In order to prove this theorem, we need the following lemma:

LEMMA 10.11. *If* $\mathbf{p}^*$ *is a fixed point of* $\mathbf{f}$, *then* $\beta(\mathbf{p}^*) = 1$.

Proof. Since $\mathbf{p}^*$ is a fixed point of $\mathbf{f}$, we have

$$\mathbf{f}(\mathbf{p}^*) = \frac{(\mathbf{p}^* + \mathbf{E}(\mathbf{p}^*))^+}{\beta(\mathbf{p}^*)} = \mathbf{p}^*. \tag{10.3}$$

This implies that, for $j = 1, 2, 3$,

$$(\mathbf{p}^* + \mathbf{E}(\mathbf{p}^*))_j^+ = \beta(\mathbf{p}^*)p_j^*. \tag{10.4}$$

Let α be an index value (equal to 1, 2, or 3) such that $p_\alpha^* > 0$. Such an α exists since the sum of the entries of $\mathbf{p}^*$ is 1. Since $\beta(\mathbf{p}^*) > 0$ (by Exercise 10.15), it follows from Equation 10.4 that

$$(\mathbf{p}^* + \mathbf{E}(\mathbf{p}^*))_\alpha^+ > 0.$$

Therefore

$$(\mathbf{p}^* + \mathbf{E}(\mathbf{p}^*))_\alpha^+ = (\mathbf{p}^* + \mathbf{E}(\mathbf{p}^*))_\alpha.$$

Plugging this back into Equation 10.4, we have

$$(\mathbf{p}^* + \mathbf{E}(\mathbf{p}^*))_{\alpha'} = \beta(\mathbf{p}^*)p_\alpha^*,$$

which yields

$$E_\alpha(\mathbf{p}^*) = (\beta(\mathbf{p}^*) - 1)p_\alpha^*.$$

Multiplying both sides by p_α^*, we obtain

$$p_\alpha^* E_\alpha(\mathbf{p}^*) = (\beta(\mathbf{p}^*) - 1)p_\alpha^* p_\alpha^*. \tag{10.5}$$

Although we derived Equation 10.5 with the assumption that $p_\alpha^* > 0$, the equation also holds when $p_\alpha^* = 0$ since it then reduces to $0 = 0$. Therefore, since p_j^* is nonnegative, Equation 10.5 holds for all p_j^*. Now, summing Equation 10.5 over all j, we obtain

$$\begin{aligned}
\mathbf{p}^* \cdot \mathbf{E}(\mathbf{p}^*) &= \sum_{j=1}^{3} p_j^* E_j(\mathbf{p}^*) \\
&= \sum_{j=1}^{3} (\beta(\mathbf{p}^*) - 1) p_j^* p_j^* \\
&= (\beta(\mathbf{p}^*) - 1) \sum_{j=1}^{3} p_j^* p_j^* \\
&= (\beta(\mathbf{p}^*) - 1) \mathbf{p}^* \cdot \mathbf{p}^*.
\end{aligned}$$

By the second form of Walras's Law (Equation 10.2), $\mathbf{p}^* \cdot \mathbf{E}(\mathbf{p}^*) = 0$. Therefore, $0 = (\beta(\mathbf{p}^*) - 1)\mathbf{p}^* \cdot \mathbf{p}^*$. Since $\mathbf{p}^* \in T$, it follows that $\mathbf{p}^* \cdot \mathbf{p}^* = 1$. Thus, it must be that $\beta(\mathbf{p}^*) - 1 = 0$, implying that $\beta(\mathbf{p}^*) = 1$, as desired. ∎

With the help of Lemma 10.11, we can now prove Theorem 10.10.

Proof of Theorem 10.10. Let $\mathbf{p}^*$ be a fixed point for $\mathbf{f}$. So $\mathbf{f}(\mathbf{p}^*) = \mathbf{p}^*$. Since $\beta(\mathbf{p}^*) = 1$, Equation 10.3 becomes

$$(\mathbf{p}^* + \mathbf{E}(\mathbf{p}^*))^+ = \mathbf{p}^*,$$

which is actually three equations of the form

$$(\mathbf{p}^* + \mathbf{E}(\mathbf{p}^*))_j^+ = p_j^*. \tag{10.6}$$

We consider two possibilities for p_j^*: either $p_j^* = 0$ or $p_j^* > 0$. Keep in mind that all prices must be nonnegative.

If $p_j^* = 0$, then $(\mathbf{p}^* + \mathbf{E}(\mathbf{p}^*))_j^+ = 0$. Therefore $E_j(\mathbf{p}^*)^+ = 0$, and it follows that $E_j(\mathbf{p}^*) \le 0$. This means that there is no excess demand for item j in this case.

Now, consider the case in which $p_j^* > 0$. Then by Equation 10.6, $(\mathbf{p}^* + \mathbf{E}(\mathbf{p}^*))_j^+ > 0$. Therefore,

$$(\mathbf{p}^* + \mathbf{E}(\mathbf{p}^*))_j^+ = (\mathbf{p}^* + \mathbf{E}(\mathbf{p}^*))_j.$$

Plugging this back into Equation 10.6, we obtain $(\mathbf{p}^* + \mathbf{E}(\mathbf{p}^*))_j = p_j^*$, and therefore $E_j(\mathbf{p}^*) = 0$.

Together, these cases imply that $E_j(\mathbf{p}^*) \le 0$ for all j. Therefore $D_j(\mathbf{p}^*) \le S_j$ for all j, and $\mathbf{p}^*$ is an equilibrium price vector. ∎

When we have an equilibrium price vector $\mathbf{p}^*$, where $D_j(\mathbf{p}^*) \le S_j$ for all j, we say the markets clear. Everyone can achieve their demand vector. The Brouwer Fixed Point Theorem tells us there is such a price. So we can all sleep well at night.

Nothing that we have said depended on the fact that we had three items available. The same would hold for an economy with hundreds of thousands of items and millions of individuals. Of course, establishing a corresponding result for a general setting would require the n-Dimensional Brouwer Fixed Point Theorem.

Exercises for Section 10.2

10.13. Describe the topological space of price vectors if the economy consists of only two items. Does the analysis of this section go through in that case?

10.14. In an economy with four goods, the resultant set of normalized price vectors will yield the analog in 4-space of the triangle T we created in 3-space. What is it? How many faces does it have, and what are their equations? How many edges and vertices does it have? What are their equations?

10.15. Show that at least one coordinate of the vector $\mathbf{p} + \mathbf{E}(\mathbf{p})$ is positive.

10.16. Suppose that our entire economy consists of only Carmen and Dexter with initial bundles $\mathbf{b}^C = (1, 1, 2)$ and $\mathbf{b}^D = (3, 4, 2)$ and initial price vector $\mathbf{p}_o = (\frac{1}{6}, \frac{1}{6}, \frac{4}{6})$. Let $\mathbf{f}$ be the price change function.

   (a) Assume that when the price vector is $\mathbf{p} = (\frac{1}{6}, \frac{3}{6}, \frac{2}{6})$, Carmen has demand vector $\mathbf{d}^C = (3, 1, 1)$ and Dexter has demand vector $\mathbf{d}^D = (8, 1, 4)$. Determine $\mathbf{f}(\mathbf{p})$ for this $\mathbf{p}$.

   (b) Assume that when the price vector is $\mathbf{p}' = (\frac{4}{6}, \frac{1}{6}, \frac{1}{6})$, Carmen has demand vector $\mathbf{d}^C = (\frac{1}{2}, \frac{1}{2}, \frac{9}{2})$ and Dexter has demand vector $\mathbf{d}^D = (2, 3, 7)$. Determine $\mathbf{f}(\mathbf{p}')$ for this $\mathbf{p}'$.

10.17. Let k be a fixed positive real number. Define

$$\mathbf{f}_k(\mathbf{p}) = \frac{(\mathbf{p} + k\mathbf{E}(\mathbf{p}))^+}{\sum_{j=1}^{3}(\mathbf{p} + k\mathbf{E}(\mathbf{p}))_j^+}.$$

Show that a fixed point $\mathbf{p}^*$ of this general price change function also yields an equilibrium price vector.

10.3 *Kakutani's Fixed Point Theorem*

In 1941, Shizuo Kakutani (1911–2004) proved a generalization of the Brouwer Fixed Point Theorem that has had powerful applications since. Instead of applying to functions from the n-ball B^n to itself, Kakutani's Fixed Point Theorem applies to so-called set-valued functions. Usually, functions associate a point x in a domain X to a point y in the range Y. Here, we will look at functions that take a point x in the domain X and send it to a nonempty subset A of the range Y. We denote such a set-valued function by $f : X \to_S Y$. (In this section we call our usual functions, whose values are points, point-valued functions, in order to distinguish them from set-valued functions.)

EXAMPLE 10.6. Consider the set of all of the people who have ever lived. Let f assign to each person the set of all people that person has ever seen. This is an example of a set-valued function. To each point (person) the function f associates a set of points (the set of people that person has ever seen).

EXAMPLE 10.7. The following are set-valued functions:

   (i) Let f assign to each real number x the set of all real numbers greater than x. We write $f(x) = \{y \in \mathbb{R} \mid y \geq x\}$.

   (ii) Let g assign to each real number x the set consisting of x and its negative. So $g(x) = \{-x, x\}$.

   (iii) Let h assign to each real number x the set $[-1, 4]$ if x is negative or the set $[-4, 1]$ if x is not negative. Therefore,

$$h(x) = \begin{cases} [-1, 4] & \text{if } x < 0, \\ [-4, 1] & \text{if } x \geq 0. \end{cases}$$

For a point-valued function $f : X \to Y$, the graph is defined to be the set $\{(x, y) \mid y = f(x)\}$ in $X \times Y$. We need a similar notion for set-valued functions:

DEFINITION 10.12. *The **graph** of a set-valued function $f : X \to_S Y$ is the subset of $X \times Y$ given by $G_f = \{(x, y) \mid y \in f(x)\}$.*

EXAMPLE 10.8. The graphs of the set-valued functions f, g, and h from Example 10.7 are shown in Figure 10.8.
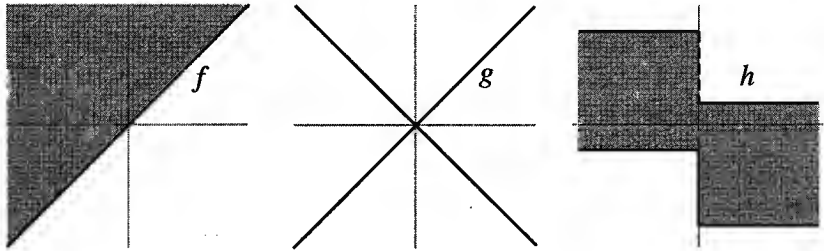
FIGURE 10.8: Graphs of the set-valued functions f, g, and h.

In the hypotheses of the Brouwer Fixed Point Theorem, the continuity of the function is critical. For set-valued functions we do not directly define continuity, but instead we work with a property closely related to it. Combining the results of Exercises 4.10 and 7.13, it follows that if Y is compact and Hausdorff, then a point-valued function $f : X \to Y$ is continuous if and only if the graph of f is a closed subset of $X \times Y$. Therefore we focus on set-valued functions having graphs that are closed sets. In Example 10.8, the set-valued functions f and g have graphs that are closed sets, but the set-valued function h does not.

Having a closed graph is advantageous since it implies that convergent sequences behave reasonably, as the following lemma indicates:

LEMMA 10.13. *Let $f : X \to_S Y$ be a set-valued function whose graph, G_f, is closed in $X \times Y$. If (x_n) is a sequence in X that converges to $x_0 \in X$, and (y_n) is a sequence in Y that converges to $y_0 \in Y$ and satisfies $y_n \in f(x_n)$ for each n, then $y_0 \in f(x_0)$.*

Proof. Form the sequence $((x, y)_n)$ in $X \times Y$ defined by $(x, y)_n = (x_n, y_n)$. Since $y_n \in f(x_n)$, this entire sequence lies in the graph of f. Since (x_n) converges to x_0 and (y_n) converges to y_0, $((x, y)_n)$ converges to (x_0, y_0). Now, either there exists N such that $(x_0, y_0) = (x_N, y_N)$, or (x_0, y_0) is distinct from every (x_n, y_n). In the first case, it directly follows that $(x_0, y_0) \in G_f$. In the second case, (x_0, y_0) must be a limit point of the set $\{(x_n, y_n)\}_{n \in \mathbb{Z}_+}$ and therefore must also be a limit point of G_f. Since G_f is closed, it follows that $(x_0, y_0) \in G_f$ in this case, as well. In either case, (x_0, y_0) is in the graph of f, and therefore $y_0 \in f(x_0)$ as we wanted to show. ∎

Now, what does it mean for a set-valued function to have a fixed point?

DEFINITION 10.14. *Given a set-valued function* $F : X \to_S X$, *a fixed point of* F *is a point* x^* *in* X *for which* $x^* \in F(x^*)$.

A fixed point of a set-valued function is a point that maps to a set containing the point, in contrast to a fixed point of a point-valued function, which is a point that simply maps to itself.

The functions considered in the Kakutani Fixed Point Theorem are assumed to have a domain that is a polyhedron in $\mathbb{R}^n$, where a polyhedron is a bounded set that can be expressed as a solution set to finitely many inequalities of the form $a_1 x_1 + \ldots + a_n x_n \le b$. (See Definition 0.12.)

The n-Dimensional Kakutani Fixed Point Theorem states that for a polyhedron X in $\mathbb{R}^n$, a set-valued function $F : X \to_S X$ has a fixed point if $F(\mathbf{x})$ is a convex subset of X for each $\mathbf{x}$ in X and if the graph of F is closed in $X \times X$.

The n-Dimensional Kakutani Fixed Point Theorem requires the n-Dimensional Brouwer Fixed Point Theorem in its proof. Here we only address the Kakutani Fixed Point Theorem in dimensions one and two, but the proof directly generalizes assuming the n-Dimensional Brouwer Fixed Point Theorem.

For the one-dimensional case, the possibilities for the polyhedron $X \subset \mathbb{R}$ are limited. In fact, such sets must be closed and bounded intervals $[a, b]$. (See Exercise 10.20.) We ask you to prove the One-Dimensional Kakutani Fixed Point Theorem in Exercise 10.21.

Now we present the Kakutani Fixed Point Theorem in dimension two:

THEOREM 10.15. *The Two-Dimensional Kakutani Fixed Point Theorem.* *Let* X *be a polyhedron in* $\mathbb{R}^2$ *and* $F : X \to_S X$ *be a set-valued function such that* $F(\mathbf{x})$ *is a convex subset of* X *for each* $\mathbf{x}$ *in* X. *If the graph of* F *is closed in* $X \times X$, *then there exists* $\mathbf{x}^* \in X$ *such that* $\mathbf{x}^* \in F(\mathbf{x}^*)$.

> *Proof.* A polyhedron in $\mathbb{R}^2$ is either a point, a line segment, or a convex polygon. The cases for a point or line segment fall under the one-dimensional version of the theorem, assigned in Exercise 10.21. Here we address the case for a convex polygon in $\mathbb{R}^2$. For simplicity, we prove the theorem when X is a triangle T in $\mathbb{R}^2$. We then discuss how the method of proof for a triangle carries over to a general convex polygon.
>
> Let T be a triangle in the plane with vertices $\mathbf{v}_0^1$, $\mathbf{v}_0^2$, and $\mathbf{v}_0^3$. (The reason for using the multiple indices will become clear as we progress.) Since T is a triangle, every point in T can be represented as a linear combination of $\mathbf{v}_0^1$, $\mathbf{v}_0^2$, and $\mathbf{v}_0^3$. Specifically, for $\mathbf{x} \in T$ we have $\mathbf{x} = \lambda_0^1 \mathbf{v}_0^1 + \lambda_0^2 \mathbf{v}_0^2 + \lambda_0^3 \mathbf{v}_0^3$, where each $\lambda_0^i \ge 0$ and $\lambda_0^1 + \lambda_0^2 + \lambda_0^3 = 1$. (See Figure 10.9.)
>
> Notice that T is homeomorphic to the disk, and therefore the Two-Dimensional Brouwer Fixed Point Theorem applies to every continuous function from T to T.
>
> To find a fixed point of F, we build a sequence $\mathbf{f}_0, \mathbf{f}_1, \mathbf{f}_2, \ldots$ of point-valued continuous functions that approximate F. By the Two-Dimensional Brouwer Fixed Point Theorem, each $\mathbf{f}_n$ has a fixed point
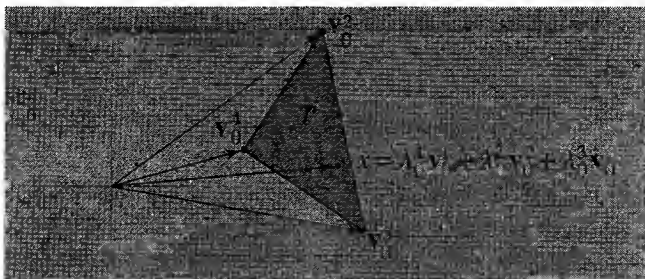
FIGURE 10.9: Every point in T is a linear combination of the vertices.

$x_n \in T$. We prove that the sequence of fixed points (x_n) has a subsequence that converges to a fixed point of F.

For each of the three vertices v_0^i, $i = 1, 2, 3$, we pick a particular point $y_0^i \in f(v_0^i)$. Then we define a function $f_0 : T \to T$ so that $f_0(v_0^i) = y_0^i$ on the vertices. We extend this linearly to the triangle by, for each $x = \lambda_0^1 v_0^1 + \lambda_0^2 v_0^2 + \lambda_0^3 v_0^3$, setting $f_0(x) = \lambda_0^1 y_0^1 + \lambda_0^2 y_0^2 + \lambda_0^3 y_0^3$.

Notice that f_0 is not a set-valued function. It is a point-valued function mapping T to itself. Moreover, because f_0 is defined to be the linear combination of its values at the vertices, it is continuous. Therefore the Two-Dimensional Brouwer Fixed Point Theorem applies, and we have a point $x_0 \in T$ such that $f_0(x_0) = x_0$.

The point x_0 is not necessarily a fixed point of the set-valued function F (but would be, for instance, if it was one of the vertices of T).

Now, to define the next function f_1 in our sequence of functions approximating F, we begin by subdividing T into four smaller triangles, as shown in Figure 10.10.
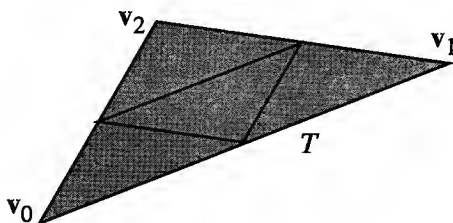
FIGURE 10.10: Subdividing T into smaller triangles.

The vertices of these four triangles are made up of the vertices of T and the midpoints of the edges of T, given by

$$\tfrac{1}{2}v_0^1 + \tfrac{1}{2}v_0^2, \ \tfrac{1}{2}v_0^1 + \tfrac{1}{2}v_0^3, \ \text{and} \ \tfrac{1}{2}v_0^2 + \tfrac{1}{2}v_0^3.$$

We define $f_1(x)$ in a manner analogous to how we defined $f_0(x)$. For each vertex v in the four triangles, we choose a point y in $F(v)$ and define $f_1(v) = y$. Then, as with f_0, we extend f_1 linearly over each triangle. Notice that if x lies in the intersection of two different triangles, then

the definition of $f_1(x)$ in terms of either triangle is the same because it depends only on the definition of f_1 on the two vertices that are the endpoints of the edge containing x.

The function f_1 maps T to itself and is continuous since it is a linear extension of the values at the vertices. Therefore, by the Brouwer Fixed Point Theorem, there is a fixed point x_1 of f_1. Here too, x_1 is not necessarily a fixed point of the set-valued function F. But it would be if it was one of the vertices of the four triangles in the subdivision of T. Let T_1 be a subdivision triangle that contains this new fixed point x_1, and assume that the vertices of T_1 are v_1^1, v_1^2, and v_1^3.

We continue this process. Specifically, assume that we have a continuous $f_{n-1} : T \to T$ with fixed point x_{n-1} in triangle $T_{n-1} \subset T$ having vertices v_{n-1}^1, v_{n-1}^2, and v_{n-1}^3.

To define f_n, we subdivide each of the triangles used in the definition of f_{n-1} into four subtriangles as described earlier. Then for each vertex v in each triangle, we define $f_n(v)$ to be a point in $F(v)$. Finally, we extend f_n linearly over each triangle in the subdivision to obtain a continuous function $f_n : T \to T$. Since f_n is continuous, it has a fixed point $x_n \in T$. Let $T_n \subset T$ be a triangle in the subdivision containing x_n, and assume T_n has vertices v_n^1, v_n^2, and v_n^3.

We now have a sequence $x_0, x_1, x_2, \ldots$ of fixed points of the functions $f_0, f_1, f_2, \ldots$. Note that $x_n = \lambda_n^1 v_n^1 + \lambda_n^2 v_n^2 + \lambda_n^3 v_n^3$ for some values λ_n^1, λ_n^2, and λ_n^3 in $[0, 1]$. Furthermore, if we have $y_n^j = f_n(v_n^j)$ for each n and j, then $x_n = f_n(x_n) = \lambda_n^1 y_n^1 + \lambda_n^2 y_n^2 + \lambda_n^3 y_n^3$ as well, yielding

$$x_n = \lambda_n^1 v_n^1 + \lambda_n^2 v_n^2 + \lambda_n^3 v_n^3 = \lambda_n^1 y_n^1 + \lambda_n^2 y_n^2 + \lambda_n^3 y_n^3. \tag{10.7}$$

Since T is a compact subset of the plane, Theorem 7.16 implies that every sequence in T has a convergent subsequence. Let x^* be the limit of a convergent subsequence (x_{j_n}) of the sequence (x_n). We show that x^* is a fixed point of the set-valued function F.

Because the side lengths of the triangles T_n shrink to zero as n approaches infinity, every sequence made up of one point from each triangle T_{j_n} must also converge to x^*. Hence, the sequence of fixed points (x_{j_n}) and the corresponding sequences of vertices $(v_{j_n}^1)$, $(v_{j_n}^2)$, and $(v_{j_n}^3)$ all converge to x^*.

Since the interval $[0, 1]$ is compact in $\mathbb{R}$, Theorem 7.16 implies that every sequence in $[0, 1]$ has a convergent subsequence. Thus the three coefficient sequences $(\lambda_{j_n}^1)$, $(\lambda_{j_n}^2)$, and $(\lambda_{j_n}^3)$ have convergent subsequences. Similarly, the three sequences $(y_{j_n}^1)$, $(y_{j_n}^2)$, and $(y_{j_n}^3)$ are sequences in the compact set T and thus also have convergent subsequences. By taking subsequences, one at a time, of all of these additional sequences, we can thereby choose a single indexing sequence (m_n) such that all ten corresponding subsequences converge: (x_{m_n}) to x^*, $(v_{m_n}^1)$ to x^*, $(v_{m_n}^2)$ to x^*, $(v_{m_n}^3)$ to x^*, $(\lambda_{m_n}^1)$ to a value λ^1, $(\lambda_{m_n}^2)$ to a value λ^2, $(\lambda_{m_n}^3)$ to a value λ^3, $(y_{m_n}^1)$ to a point y^1, $(y_{m_n}^2)$ to a point y^2, and $(y_{m_n}^3)$ to a point y^3.

As m_n approaches infinity, we see from Equation 10.7 that $\mathbf{x}^* = \lambda^1\mathbf{y}^1 + \lambda^2\mathbf{y}^2 + \lambda^3\mathbf{y}^3$. Furthermore $\lambda^1 + \lambda^2 + \lambda^3 = 1$, and $\lambda^1, \lambda^2, \lambda^3 \in [0, 1]$. Therefore $\mathbf{x}^*$ is in the "triangle" with vertices $\mathbf{y}^1, \mathbf{y}^2$, and $\mathbf{y}^3$. (We use quotes since the triangle could be a line segment or a point if exactly two of the $\mathbf{y}^i$ are equal or if they all are equal, respectively.)

Since F has a closed graph in $T \times T$, Lemma 10.13 applies. Therefore, since the sequences $(\mathbf{v}_{m_n}^1)$ and $(\mathbf{y}_{m_n}^1)$ converge to $\mathbf{x}^*$ and $\mathbf{y}^1$, respectively, and since $\mathbf{y}_{m_n}^1 \in F(\mathbf{v}_{m_n}^1)$ for each m_n, we know that $\mathbf{y}^1 \in F(\mathbf{x}^*)$. Similarly, $\mathbf{y}^2 \in F(\mathbf{x}^*)$ and $\mathbf{y}^3 \in F(\mathbf{x}^*)$.

But $\mathbf{x}^* = \lambda^1\mathbf{y}^1 + \lambda^2\mathbf{y}^2 + \lambda^3\mathbf{y}^3$ is in the triangle with vertices $\mathbf{y}^1, \mathbf{y}^2$, and $\mathbf{y}^3$, all of which lie in $F(\mathbf{x}^*)$. Since $F(\mathbf{x}^*)$ is convex, $\mathbf{x}^*$ must also be contained in $F(\mathbf{x}^*)$. In other words, $\mathbf{x}^* \in F(\mathbf{x}^*)$, as we wished to show.

We have now proven the Two-Dimensional Kakutani Fixed Point Theorem assuming that the domain is a triangle in the plane. To prove the result for a general convex polygon in the plane, we use the same approach, but we start by subdividing the polygon into triangles. The initial approximating function $\mathbf{f}_0$ is first defined on the vertices of these triangles and then is extended linearly to each of the triangles, just as was done in the foregoing process. Then, as was done previously, successive approximating functions $\mathbf{f}_n$ are defined by subdividing each triangle considered at the previous stage, defining $\mathbf{f}_n$ on the vertices of the new triangles, and extending the definition linearly to each triangle. The same argument yields a point $\mathbf{x}^*$ such that $\mathbf{x}^* \in F(\mathbf{x}^*)$. ∎

This method of proving the Two-Dimensional Kakutani Fixed Point Theorem carries over to the general n-dimensional version, with the n-Dimensional Brouwer Fixed Point Theorem required along the way.

Kakutani's Fixed Point Theorem contains the Brouwer Fixed Point Theorem as a special case, but we need to view the domain of the Brouwer Fixed Point Theorem as a polyhedron to see this. Let

$$P^n = \{(x_1, \ldots, x_n) \in \mathbb{R}^n \mid -1 \le x_j \le 1 \text{ for } j = 1, \ldots, n\}.$$

The set P^n is an n-dimensional cube. By definition P^n is a polyhedron, and it is homeomorphic to the n-ball, B^n. Let us consider the Brouwer Fixed Point Theorem on P^n. If $\mathbf{f} : P^n \to P^n$ is a continuous point-valued function, then $\mathbf{f}$ can be considered a set-valued function mapping $\mathbf{x} \in P^n$ to the single-point set $\{\mathbf{f}(\mathbf{x})\}$. The domain of $\mathbf{f}$ is a compact polyhedron, and $\{\mathbf{f}(\mathbf{x})\}$ is a single-point set in P^n and therefore is a nonempty convex subset of the domain. Furthermore, the point-valued function $\mathbf{f}$ is continuous; thus the graph of $\mathbf{f}$ is closed in $P^n \times P^n$ by Exercise 4.10. Kakutani's Fixed Point Theorem then applies, implying that there exists $\mathbf{x} \in P^n$ such that $\mathbf{x} \in \{\mathbf{f}(\mathbf{x})\}$, and therefore $\mathbf{x} = \mathbf{f}(\mathbf{x})$.

It is important to realize, however, that although the Brouwer Fixed Point Theorem is a special case of the Kakutani Fixed Point Theorem, the latter does not supplant the former, since the Brouwer Fixed Point Theorem is needed in the proof of the Kakutani Fixed Point Theorem.

Exercises for Section 10.3

10.18. For each of the following set-valued functions, draw the graph and determine whether or not $F(x)$ is convex for all values of x in the domain:

(a) Let $F : \mathbb{R} \to_S \mathbb{R}$ be given by $F(x) = \{y \in \mathbb{R} \mid y \leq x\}$.

(b) Let $G : \mathbb{R} \to_S \mathbb{R}$ be given by $G(x) = \{nx \in \mathbb{R} \mid n \in \mathbb{Z}\}$.

(c) Let $H : \mathbb{R} \to_S \mathbb{R}^2$ be given by $H(x) = \{(y_1, y_2) \in \mathbb{R}^2 \mid y_1 y_2 = x\}$.

10.19. Define a relation on $\mathbb{R}$ by $a \sim b$ if there exists an integer c such that $a, b \in [c, c+1)$.

(a) Show that $\sim$ is an equivalence relation.

(b) Let $f : \mathbb{R} \to_S \mathbb{R}$, be the set-valued function defined by $f(x) = [x]$, where $[x]$ is the equivalence class of x under the equivalence relation $\sim$. Sketch the graph of f.

(c) With f as defined in (b), is $f(x)$ convex for each $x \in \mathbb{R}$?

(In general an equivalence relation on a set X gives rise to a natural set-valued function, sending each point in X to the equivalence class of points in X that contains it.)

10.20. Prove that a polyhedron in $\mathbb{R}$ must be a closed and bounded interval $[a, b]$.

10.21. Explain the proof of the Kakutani Fixed Point Theorem in the case that X is the interval $[0, 1]$. Include pictures of n graphs of F, f_0, and f_1 and the fixed points x_0 and x_1 in your explanation.

10.22. Determine whether or not the Kakutani Fixed Point Theorem applies to each of the following set-valued functions. Note that you need to check that X is a polyhedron, that $F(x)$ is convex for each $x \in X$, and that the graph of F is closed. If Kakutani's Fixed Point Theorem applies, find as many fixed points as you can for F.

(a) $X = [0, 1]$ and $F(x) = \{y \in [0, 1] \mid y \geq 1 - x\}$.

(b) $X = [0, 1]$ and $F(x) = \{y \in [0, 1] \mid \frac{y}{x}$ is irrational$\}$.

(c) $X = [0, 1]$ and $F(x) = \{y \in [0, 1] \mid y \leq x^2\}$.

(d) $X = [0, \infty)$ and $F(x) = \{y \in [0, \infty) \mid y \geq 2x\}$.

(e) $X = [0, 1] \cup [2, 3]$ and

$$F(x) = \begin{cases} [2, 3] & \text{if } x \in [0, 1], \\ [0, 1] & \text{if } x \in [2, 3]. \end{cases}$$

(f) $X = [0, 1] \times [0, 1]$ and

$$F(x, y) = \{(x, y) \in [0, 1] \times [0, 1] \mid x \geq y \text{ and } y = 1 - x\}.$$

10.23. For each of the following hypotheses in the Kakutani Fixed Point Theorem, give an example to show that the theorem does not hold if the hypothesis is dropped. That is, find a set-valued function $F : X \to_S X$ that satisfies each of the other two hypotheses, but does not have a fixed point.

(a) X is a polyhedron.

(b) $F(x)$ is convex for each $x \in X$.

(c) G_F is closed.

10.4   *Game Theory and the Nash Equilibrium*

In this section we introduce game theory and provide an application of the Kakutani Fixed Point Theorem, using it to prove John Nash's celebrated theorem on the existence of equilibria in n-person games. Although all of what follows applies to n-person games, we will restrict ourselves to at most three players in order to keep the notation straightforward.

Suppose that Elaine, George, and Newman agree to play a game. At a given turn, each of the three has a finite number of moves that she or he can make. For Elaine, these choices are labeled 1 through n_E. For George, they are labeled 1 through n_G. And for Newman, they are labeled 1 through n_N. These moves are called pure strategies.

We set the rules of our game so that all three players make a move without knowing what the other players have chosen to do. Then they each receive some payoff. When Elaine makes move i, George makes move j, and Newman makes move k. Consequently, Elaine receives payoff E_{ijk}, George receives payoff G_{ijk}, and Newman receives payoff N_{ijk}. Each player has an associated $n_E \times n_G \times n_N$ three-dimensional payoff array and all three are familiar with the three payoff arrays.

Of course, if Elaine always made the same move, it would not take George and Newman long to decide how best to maximize their own payoffs relative to hers. So instead of consistently making move i, Elaine may choose to play a different strategy. She may choose to play each move with a certain probability.

It makes good sense to do this. For instance, in a poker game, a player does not want to bluff every time certain cards are in her hand, as the other players may soon pick up on the fact that she consistently does this. Rather, she should bluff with a certain probability that she has predetermined. This will be her strategy. Then no particular card pattern will be associated with her bluffing.

We denote the probability that Elaine plays move i by p_i, that George plays move j by q_j, and that Newman plays move k by r_k. We make the basic probability assumptions that

$$p_i \geq 0, \quad q_j \geq 0, \quad \text{and } r_k \geq 0 \text{ for all } i, j, k, \tag{10.8}$$

and

$$\sum_i p_i = 1, \quad \sum_j q_j = 1, \quad \text{and } \sum_k r_k = 1. \tag{10.9}$$

DEFINITION 10.16. *The corresponding probability vectors*

$$\mathbf{p} = (p_1, p_2, \ldots, p_{n_E}), \mathbf{q} = (q_1, q_2, \ldots, q_{n_G}),$$
$$\text{and } \mathbf{r} = (r_1, r_2, \ldots, r_{n_N})$$

*are called **mixed strategies**.*

Now, the payoff for each player becomes an expected value. For Elaine, it is given by

$$E(\mathbf{p}, \mathbf{q}, \mathbf{r}) = \sum_{i,j,k} E_{ijk} p_i q_j r_k.$$

This expected value is the sum of the possible payoffs multiplied by the probability that such a payoff occurs. We similarly define the expected values $G(\mathbf{p}, \mathbf{q}, \mathbf{r})$ and $N(\mathbf{p}, \mathbf{q}, \mathbf{r})$ for George and Newman, respectively. If Elaine, George, and Newman play their games enough times, we expect their payoffs to average out to approximately $E(\mathbf{p}, \mathbf{q}, \mathbf{r})$, $G(\mathbf{p}, \mathbf{q}, \mathbf{r})$, and $N(\mathbf{p}, \mathbf{q}, \mathbf{r})$, respectively.

EXAMPLE 10.9. Suppose Elaine, George, and Newman are deciding where to go for dinner. They can choose either the Happy Star Chinese Restaurant or the New Yorker Diner. All three agree to simultaneously yell out either "Chinese" or "diner." If two of them agree and the third does not, they all go to the restaurant chosen by the two who agreed, and the third has to pay 10 dollars to each of the others for their dinner. If all three agree, they go to the restaurant that they all chose, and everyone pays for their own dinner.

Elaine's payoff array then has entries of the form

$$E_{ijk} = \begin{cases} -20 & \text{if } i \neq j = k, \\ 0 & \text{if } i = j = k, \\ 10 & \text{otherwise.} \end{cases}$$

The payoff arrays for George and Newman are defined similarly.

Let us assume that Elaine, George, and Newman are equally likely to pick the Chinese restaurant or to pick the diner on any given evening. Then their mixed strategies are all the same, namely $\mathbf{p} = \mathbf{q} = \mathbf{r} = (\frac{1}{2}, \frac{1}{2})$. From this, we can compute the expected value of Elaine's payoff. We find that

$$E = -20(\tfrac{1}{4}) + 0(\tfrac{1}{4}) + 10(\tfrac{1}{2}) = 0.$$

Similarly, the expected values for George and Newman are also both 0.

Now, suppose that in a particular game, Elaine knows that George and Newman are using mixed strategies $\mathbf{q}$ and $\mathbf{r}$, respectively. Elaine's goal is to obtain the largest possible payoff; that is, she wants to maximize her expected value, given $\mathbf{q}$ and $\mathbf{r}$. Thus, she wants to find a strategy $\mathbf{p}$ so that $E(\mathbf{p}, \mathbf{q}, \mathbf{r}) \geq E(\mathbf{p}', \mathbf{q}, \mathbf{r})$ over all possible probability vectors $\mathbf{p}'$. There may be more than one such strategy for Elaine, and therefore we have the following definition:

DEFINITION 10.17. *Let $P(\mathbf{q}, \mathbf{r})$ be the collection of all probability vectors $\mathbf{p}$ that satisfy $E(\mathbf{p}, \mathbf{q}, \mathbf{r}) \geq E(\mathbf{p}', \mathbf{q}, \mathbf{r})$ over all mixed strategies $\mathbf{p}'$. This is called the **set of optimal mixed strategies associated to mixed strategies** $\mathbf{q}$ **and** $\mathbf{r}$.*

Similarly, $Q(\mathbf{p}, \mathbf{r})$ is the set of optimal mixed strategies for George, maximizing his payoff when $\mathbf{p}$ and $\mathbf{r}$ are given, and $R(\mathbf{p}, \mathbf{q})$ is the corresponding set of optimal mixed strategies for Newman, given $\mathbf{p}$ and $\mathbf{q}$.

Now, how does Elaine choose vectors to maximize the expected value $E(\mathbf{p}, \mathbf{q}, \mathbf{r})$? For a fixed $\mathbf{q}$ and $\mathbf{r}$, the expected value is a linear equation in the components of $\mathbf{p}$, as follows:

$$E(\mathbf{p}, \mathbf{q}, \mathbf{r}) = \sum_{i,j,k} E_{ijk} p_i q_j r_k = \sum_i (\sum_{j,k} E_{ijk} q_j r_k) p_i. \qquad (10.10)$$

Letting $a_i = \sum_{j,k} E_{ijk} q_j r_k$, we can rewrite Equation 10.10 as

$$E(\mathbf{p}, \mathbf{q}, \mathbf{r}) = a_1 p_1 + a_2 p_2 + \ldots + a_{n_E} p_{n_E}.$$

Once we have rewritten the expected value in this way, it becomes clear how to maximize it, as in the following example:

EXAMPLE 10.10. Suppose that in a particular game played by Elaine, George, and Newman, there are five possible moves from which Elaine can choose. Further suppose that for the fixed mixed strategies $\mathbf{q}$ and $\mathbf{r}$ of George and Newman, respectively, The expected value of Elaine's payoff is

$$E(\mathbf{p}, \mathbf{q}, \mathbf{r}) = 3p_1 + 5p_2 + 4p_3 + 5p_4 + p_5.$$

She wants to find $\mathbf{p} = (p_1, p_2, p_3, p_4, p_5)$ to maximize $E(\mathbf{p}, \mathbf{q}, \mathbf{r})$. Clearly, the best choice for Elaine is to make the components of $\mathbf{p}$ with smaller coefficients as tiny as possible and those with the largest coefficients as big as possible. In this case, the largest coefficient is 5. So p_1, p_3, and p_5 should be chosen to be 0. Since the coefficients of p_2 and p_4 are both 5, she could choose $p_2 = 1$ and $p_4 = 0$. Or she could choose $p_2 = 0$ and $p_4 = 1$. In fact, she could choose p_2 and p_4 to be any combination that adds to 1. She will still have a probability vector that maximizes the payoff. Therefore she chooses her set of optimal mixed strategies to be

$$P(\mathbf{q}, \mathbf{r}) = \{(0, p_2, 0, p_4, 0) \mid p_2 + p_4 = 1, p_2 = 0, p_4 = 0\}.$$

In general, we see that the set of optimal mixed strategies $P(\mathbf{q}, \mathbf{r})$ is given by

$$P(\mathbf{q}, \mathbf{r}) = \{\mathbf{p} \mid \sum_i p_i = 1 \text{ and } p_i = 0 \text{ if } a_i < \max_j\{a_j\}\}. \qquad (10.11)$$

It follows that $P(\mathbf{q}, \mathbf{r})$ is a closed, bounded, and convex subset of $\mathbb{R}^{n_E}$. (See Exercise 10.26.)

Now, if Elaine chooses an optimal mixed strategy $\mathbf{p}$ in response to the mixed strategies $\mathbf{q}$ and $\mathbf{r}$ of George and Newman, then George might want to change his mixed strategy to an optimal one, given that Elaine and Newman have chosen mixed strategies $\mathbf{p}$ and $\mathbf{r}$. And, of course, Newman might want to change his mixed strategy as well.

We would like to know if there is a choice of mixed strategies for all of the players such that every player is using an optimal mixed strategy at the same time. Under that circumstance, there would be no incentive for the players to change strategies.

DEFINITION 10.18. *Mixed strategy vectors* **p**, **q**, *and* **r** *are said to **solve the game** if* **p** $\in$ *P*(**q**, **r**), **q** $\in$ *Q*(**p**, **r**), *and* **r** $\in$ *R*(**p**, **q**). *We then say that* **p**, **q**, *and* **r** *are a **Nash equilibrium**.*

In a Nash equilibrium, no player has an incentive to change his or her strategy, assuming no one else changes theirs. Each player is making as much as he or she possibly can, given the current strategies of the other players.

In his doctoral dissertation at Princeton University in 1950, John Nash proved that every game has at least one equilibrium (which later became known as a Nash equilibrium). His results revolutionized the field of game theory and subsequently had a significant impact on economics and the social sciences. In 1994, the Nobel Prize in economics was awarded to John Nash, Reinhard Selten, and John C. Harsanyi, with the latter two receiving the award for their refinements to the Nash equilibrium theory for situations where multiple equilibria exist and situations where each player does not have complete information about the other players' strategies in a game.

We present and prove Nash's Theorem momentarily, but first we look at an example.

EXAMPLE 10.11. Let us determine the Nash equilibria for the restaurant game of Example 10.9. We consider the situation from Elaine's point of view. Let $p = (a, 1 - a)$ be her mixed strategy. So a is the probability that she picks the Chinese restaurant, and $1 - a$ is the probability that she picks the diner. Similarly, let George's mixed strategy be $q = (b, 1 - b)$, and let Newman's be $r = (c, 1 - c)$.

Then the expected value for Elaine's payoff is

$$E = 10[ab(1 - c) + a(1 - b)c + (1 - a)(1 - b)c + (1 - a)b(1 - c)]$$
$$-20[a(1 - b)(1 - c) + (1 - a)(bc)],$$

which simplifies to

$$E = 10[2a(b + c - 1) + c + b - 3cb].$$

For the time being, assume that b and c are fixed. Let us see how Elaine can maximize E by choosing an appropriate a. We consider three different possibilities for the sum $b + c$. First, if $b + c > 1$, then Elaine obtains maximum E by choosing $a = 1$. Note that with $a = 1$, both $a + b$ and $a + c$ are greater than 1, since both b and c must be nonzero for $b + c > 1$ to hold. Then, with $a + c > 1$, by the same argument we used for Elaine, it follows that George can maximize his expected value by choosing $b = 1$. Similarly, by choosing $c = 1$, Newman can maximize his expected value. So $a = b = c = 1$ is a Nash equilibrium. This equilibrium corresponds to having everyone always choose the Chinese restaurant.

In like fashion, it can be shown that if $b + c < 1$, then all three players will maximize their expected values with the choices $a = b = c = 0$. (See Exercise 10.27.) Therefore everyone always choosing the diner is a second Nash equilibrium.

The last possibility to consider is $b + c = 1$. In this case it does not matter what probability Elaine picks, as far as her expected value goes. But from the perspective of George and Newman, it does matter. If either $a + b$ or $a + c$ does not equal 1, then, as we saw earlier, the corresponding player will change his probability to 0 or 1, forcing the other players to change, and having everyone settle on one of the two equilibria already mentioned. The only case where this does not occur is if $a + b = 1$, $a + c = 1$, and $b + c = 1$. In this case we have $a = \frac{1}{2}$, $b = \frac{1}{2}$, and $c = \frac{1}{2}$. So this is a third Nash equilibrium; it corresponds to having each player choose each restaurant half of the time.

Before we proceed with Nash's Theorem, we review the components that lead up to it:

(i) In an n-person game each player has a choice of moves to make, and each move is made without knowing the moves of the others.

(ii) Associated to each player is a payoff array giving the payoffs the player receives for all of the possible outcomes associated with each player making a choice of move.

(iii) Each player can choose a mixed strategy, and the combined choices of mixed strategies determine the average payoff, or expected value, for each player.

(iv) Each player has a set of optimal mixed strategies for each choice of mixed strategies by the other players. Each optimal mixed strategy results in the maximum possible expected value for the player given the mixed strategies of the others.

(v) A Nash equilibrium is a choice of mixed strategies that yields, for each player, the maximum possible expected value relative to the mixed strategies of the other players.

THEOREM 10.19. Nash's Theorem.   *There exists a Nash equilibrium for every n-person game.*

Proof. We present the proof for three-person games to keep the notation simple, but the same proof applies to n-person games. Given the three payoff arrays E_{ijk}, G_{ijk}, and N_{ijk}, we need to demonstrate that there is a set of mixed strategies $\mathbf{p}^*$, $\mathbf{q}^*$, and $\mathbf{r}^*$ that solves the game.

Let $m = n_E + n_G + n_N$. For a choice of mixed strategy vectors $\mathbf{p} = (p_1, \dots, p_{n_E})$, $\mathbf{q} = (q_1, \dots, q_{n_G})$, and $\mathbf{r} = (r_1, \dots, r_{n_N})$, we define an m-vector by concatenating the components of these vectors, as follows:

$$\mathbf{w} = (\mathbf{p}, \mathbf{q}, \mathbf{r}) = (p_1, \dots, p_{n_E}, q_1, \dots, q_{n_G}, r_1, \dots, r_{n_N}).$$

Each such vector $\mathbf{w}$ represents a combined choice of individual mixed strategies from each player. Its components must satisfy Inequalities 10.8 and Equations 10.9. It follows that the set of possible vectors $\mathbf{w}$ is a polyhedron X in $\mathbb{R}^m$.

We define a set-valued function on X by

$$F(\mathbf{p}, \mathbf{q}, \mathbf{r}) = \{(\mathbf{p}', \mathbf{q}', \mathbf{r}') \mid \mathbf{p}' \in P(\mathbf{q}, \mathbf{r}), \mathbf{q}' \in Q(\mathbf{p}, \mathbf{r}), \mathbf{r}' \in R(\mathbf{p}, \mathbf{q})\}.$$

In other words, the vector $\mathbf{w} = (\mathbf{p}, \mathbf{q}, \mathbf{r})$ is sent to the collection of vectors that have the property that their first n_E components are an optimal strategy for Elaine, given that George and Newman stick with strategies $\mathbf{q}$ and $\mathbf{r}$; their second n_G components are an optimal strategy for George, given that Elaine and Newman stick with strategies $\mathbf{p}$ and $\mathbf{r}$; and their last n_N components are an optimal strategy for Newman, given that Elaine and George stick with strategies $\mathbf{p}$ and $\mathbf{q}$.

If we show that there is a $\mathbf{w}^*$ such that $\mathbf{w}^* \in F(\mathbf{w}^*)$, then we will have proven the theorem, because such a vector $\mathbf{w}^* = (\mathbf{p}^*, \mathbf{q}^*, \mathbf{r}^*)$ is made up of three vectors, $\mathbf{p}^*, \mathbf{q}^*$, and $\mathbf{r}^*$ such that $\mathbf{p}^* \in P(\mathbf{q}^*, \mathbf{r}^*)$, $\mathbf{q}^* \in Q(\mathbf{p}^*, \mathbf{r}^*)$, and $\mathbf{r}^* \in R(\mathbf{p}^*, \mathbf{q}^*)$. Thus, we need to prove that there is a fixed point for the set-valued function $F : X \to_S X$.

We show that Kakutani's Fixed Point Theorem applies to F. It was previously observed that X is a polyhedron in $\mathbb{R}^m$. Since each of $P(\mathbf{q}, \mathbf{r})$, $Q(\mathbf{p}, \mathbf{r})$, and $R(\mathbf{p}, \mathbf{q})$ is convex, it must be that $F(\mathbf{w})$ is convex as well. (See Exercise 10.25.) Hence, we need only show that the graph G_F is a closed subset of $X \times X \subset \mathbb{R}^{2m}$. Let $(\mathbf{x}_0, \mathbf{y}_0)$ be a limit point of G_F. For each positive integer i, pick a point $(\mathbf{x}_i, \mathbf{y}_i)$ in the intersection of G_F with the ball of radius $1/i$ centered at $(\mathbf{x}_0, \mathbf{y}_0)$. We obtain a sequence of points $(\mathbf{x}_i, \mathbf{y}_i)$ in G_F converging to $(\mathbf{x}_0, \mathbf{y}_0)$. Let $\mathbf{x}_i = (\mathbf{p}_i, \mathbf{q}_i, \mathbf{r}_i)$, $\mathbf{y}_i = (\mathbf{s}_i, \mathbf{t}_i, \mathbf{u}_i)$, $\mathbf{x}_0 = (\mathbf{p}_0, \mathbf{q}_0, \mathbf{r}_0)$, and $\mathbf{y}_0 = (\mathbf{s}_0, \mathbf{t}_0, \mathbf{u}_0)$. We then have the following convergent sequences: $(\mathbf{p}_i)$ to $\mathbf{p}_0$, $(\mathbf{q}_i)$ to $\mathbf{q}_0$, $(\mathbf{r}_i)$ to $\mathbf{r}_0$, $(\mathbf{s}_i)$ to $\mathbf{s}_0$, $(\mathbf{t}_i)$ to $\mathbf{t}_0$, and $(\mathbf{u}_i)$ to $\mathbf{u}_0$.

Note that $\mathbf{y}_i \in F(\mathbf{x}_i)$ for all i. Therefore $\mathbf{s}_i \in P(\mathbf{q}_i, \mathbf{r}_i)$, $\mathbf{t}_i \in Q(\mathbf{p}_i, \mathbf{r}_i)$, and $\mathbf{u}_i \in R(\mathbf{p}_i, \mathbf{q}_i)$. Hence, for all $\mathbf{p}'$, $\mathbf{q}'$, and $\mathbf{r}'$, we have

$$E(\mathbf{s}_i, \mathbf{q}_i, \mathbf{r}_i) \geq E(\mathbf{p}', \mathbf{q}_i, \mathbf{r}_i),$$

$$G(\mathbf{p}_i, \mathbf{t}_i, \mathbf{r}_i) \geq G(\mathbf{p}_i, \mathbf{q}', \mathbf{r}_i), \text{ and}$$

$$N(\mathbf{p}_i, \mathbf{q}_i, \mathbf{u}_i) \geq N(\mathbf{p}_i, \mathbf{q}_i, \mathbf{r}').$$

Since E, G, and N are continuous functions, these inequalities hold as we take the limit as i approaches infinity. (See Exercise 10.24.) This implies that

$$E(\mathbf{s}_0, \mathbf{q}_0, \mathbf{r}_0) \geq E(\mathbf{p}', \mathbf{q}_0, \mathbf{r}_0),$$

$$G(\mathbf{p}_0, \mathbf{t}_0, \mathbf{r}_0) \geq G(\mathbf{p}_0, \mathbf{q}', \mathbf{r}_0), \text{ and}$$

$$N(\mathbf{p}_0, \mathbf{q}_0, \mathbf{u}_0) \geq N(\mathbf{p}_0, \mathbf{q}_0, \mathbf{r}').$$

Therefore $\mathbf{s}_0 \in P(\mathbf{q}_0, \mathbf{r}_0)$, $\mathbf{t}_0 \in Q(\mathbf{p}_0, \mathbf{r}_0)$, and $\mathbf{u}_0 \in R(\mathbf{p}_0, \mathbf{q}_0)$. Thus, $\mathbf{y}_0 \in F(\mathbf{x}_0)$, implying that $(\mathbf{x}_0, \mathbf{y}_0)$ is in the graph of F. It follows that the graph of F is closed. The Kakutani Fixed Point Theorem now applies, and therefore there exists a $\mathbf{w}^*$ such that $\mathbf{w}^* \in F(\mathbf{w}^*)$, as we wished to show. ∎

Thus, for our three-player game, we are assured that a Nash equilibrium exists, and therefore it is possible for each player to choose a mixed strategy that results in maximum expected value relative to the others' choices of mixed strategy.

Even though the previous proof involves a three-person game, we need the general version of the Kakutani Fixed Point Theorem to establish the existence of the Nash equilibrium. This is because it is the total number of plays that the players can make, rather than the total number of players, that determines the dimension of the space in which we are working.

As we indicated earlier, these results carry over to n-person games for any positive integer n. The proof for the general case is essentially the same as the proof presented here, with the Kakutani Fixed Point Theorem establishing the existence of the Nash equilibrium.

EXAMPLE 10.12. Here we look at an example of the classic prisoner's dilemma two-person game. George and Newman have been charged with a crime, and each of them is interviewed separately about it. Since the prosecutor does not have sufficient evidence for a conviction, George and Newman are each offered a minimal sentence for telling on the other. So, George and Newman can each either remain silent or tell on the other.

If both George and Newman choose to be silent, each will receive a six-month sentence for a lesser crime. If they each tell on the other, then they each receive a reduced 24-month sentence because they assisted in the prosecution. If one tells while the other is silent, then the one who told receives a three-month sentence while the one who remained silent receives the full 72-month sentence.

For each player, the four possible outcomes and the resulting consequences are displayed in the payoff matrices in Figure 10.11. The payoffs $G_{i,j}$ and $N_{i,j}$ in the matrices represent the sentences (in months) received by George and Newman, respectively, in each circumstance. Here, of course, the desire of each is to minimize their sentence.

FIGURE 10.11: The prisoner's dilemma payoff matrices.

It is straightforward to show that the only Nash equilibrium in this game corresponds to having both players always tell on the other. (See Exercise 10.28.) Note, however, that the Nash equilibrium does not yield the best cooperative outcome of the game. If George and Newman agreed in advance to be silent,

they would receive a better result than the one in the Nash equilibrium. However, since the game is not played cooperatively, for each player the threat of having the other player tell on them while they remain silent leads the player away from silence into the equilibrium.

Exercises for Section 10.4

10.24. Assume that $f : X \to \mathbb{R}$ is a continuous function, (x_i) converges to x_0, and (y_i) converges to y_0. Show that if $f(x_i) \geq f(y_i)$ for all i, then $f(x_0) \geq f(y_0)$.

10.25. Prove that if A is a convex subset of $\mathbb{R}^n$ and B is a convex subset of $\mathbb{R}^m$, then $A \times B$ is a convex subset of $\mathbb{R}^{n+m}$.

10.26. Given that the set of optimal mixed strategies $P(\mathbf{q}, \mathbf{r})$ satisfies Equation 10.11, prove that $P(\mathbf{q}, \mathbf{r})$ is a closed, bounded, and convex subset of $\mathbb{R}^{n_E}$.

10.27. Show that in Example 10.11, if we have a Nash equilibrium with $b + c < 1$, then $a = b = c = 0$.

10.28. Show that the only Nash equilibrium in the prisoner's dilemma game in Example 10.12 corresponds to having both players always tell on the other.

10.29. Elaine, George, and Newman are playing the same restaurant game as in Example 10.9. Suppose, however, that they have payoffs as follows:

    (i) If George and Newman agree, regardless of what Elaine calls out, they then go to the restaurant that George and Newman chose, and Elaine has to pay them three dollars each for coffee.

    (ii) If Elaine agrees with either George or Newman, but not both, they then go to the restaurant chosen by the two who agreed, and the third person pays 10 dollars to each of the others for dinner.

    (iii) If all three agree, they then go to the restaurant that they all chose, and George and Newman each pay Elaine five dollars toward her dinner.

Find all Nash equilibria for this game.

10.30. Elaine, George, and Newman can choose between the Happy Star Chinese Restaurant, Bobo's Burgers, or Pat's Pizza in the same game as in Example 10.9. Suppose that when two or three of them agree, the payout is the same as in the game in Example 10.9, and when all three disagree they all eat at home and there is no payout.

    (a) If all three use a mixed strategy vector $(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$ to determine which restaurant to yell out, what are the resulting expected values?

    (b) Find all Nash equilibria for this game.

Chapter 11

Embeddings

Along with the study of topological spaces, an important component of topology is understanding how topological spaces can sit within each other. Recall from Section 4.2 that we define an embedding of one topological space X in another topological space Y to be a homeomorphism from X to a subspace of Y. In this way, we make precise the idea of a copy of X sitting in Y.

In this chapter we explore a variety of important embedding theorems in topology, starting with an overview of a number of interesting embedding questions and results in Section 11.1. In Section 11.2 we prove the Jordan Curve Theorem, a classic theorem in topology, which says that an embedded circle in the plane separates the plane into two components, each of which has the embedded circle as its boundary. In Section 11.3 we introduce a digital version of the Jordan Curve Theorem, and we discuss an application of it to digital image processing.

11.1 *Some Embedding Results*

In this section we consider two fundamental embedding questions:

(i) Given topological spaces X and Y, is there an embedding of X in Y?

(ii) If there is an embedding of X in Y, what can we say about how the image of the embedding of X sits in Y?

DEFINITION 11.1. *If $f : X \to Y$ is an embedding of space X into space Y we say that X is **embeddable** in Y, or that we can **embed** X in Y. Furthermore, we call Y the **embedding space** and we call the image $f(X)$ the **embedded image of** X.*

We introduced a result related to the first of these questions earlier in the book, without expressing it in terms of embeddings. When we introduced the Klein bottle in Section 3.4, we indicated that it does not exist in 3-space. That is to say, there is no embedding of the Klein bottle in $\mathbb{R}^3$. The same is true for the projective plane. We do not develop the tools to prove these nonembedding results in this book, but in this section we demonstrate that the Klein bottle is embeddable in $\mathbb{R}^4$, and we ask you to find an embedding of the projective plane in $\mathbb{R}^4$ in Exercise 11.9.

Let us begin by considering embeddings of the circle into topological spaces. Recall that the image of such an embedding is called a simple closed curve. Since S^1 is compact, it follows that simple closed curves are compact subsets of the embedding space.

There are no simple closed curves in $\mathbb{R}$, and the only simple closed curve in the circle S^1 is the circle itself. (See Exercise 11.1.)

We need another dimension in our embedding space for simple closed curves to be interesting, so let us consider simple closed curves in the plane. They can be straightforward, as we see on the left in Figure 11.1, or they can be quite complicated, winding throughout a region in the plane, as we see on the right. Despite this, simple closed curves in the plane have much in common, as asserted by the Jordan Curve Theorem.
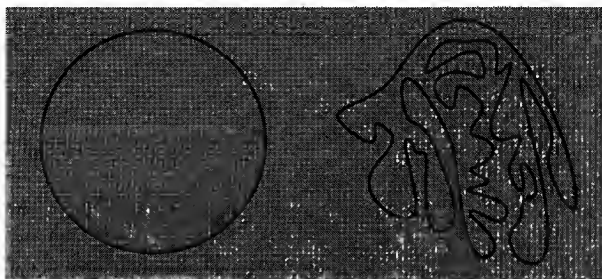
FIGURE 11.1: Simple closed curves in the plane.

THEOREM 11.2. *The Jordan Curve Theorem..* *Let S be a simple closed curve in $\mathbb{R}^2$. Then $\mathbb{R}^2 - S$ consists of two components, and S is the boundary in $\mathbb{R}^2$ of each of the components.*

This result was first presented by Camille Jordan (1838–1922) in 1887 in his work *Cours d'Analyse de l'École Polytechnique*. However, the proof he gave was incorrect. It was not until eighteen years later, in 1905, that Oswald Veblen (1880–1960) provided the first correct proof in [Veb]. We prove the Jordan Curve Theorem in the next section.

A simple closed curve in the plane is compact and therefore is closed and bounded. Since it is bounded, its complement in the plane has one unbounded component. (See Exercise 6.19.) Thus it follows from the Jordan Curve Theorem that the complement of a simple closed curve in the plane has one unbounded component and one bounded component.

DEFINITION 11.3. *Given a simple closed curve S in the plane, we call the bounded component of $\mathbb{R}^2 - S$ the **inside** of S and the unbounded component the **outside** of S.*

It is an interesting consequence of the Jordan Curve Theorem that if we have two points in the complement of a simple closed curve S in the plane and we draw a line L from one point to the other, then if L crosses S an even number of times, the points are in the same component of the complement of S, and if L crosses S an odd number of times, the points are in different components of the complement of S. (See Figure 11.2.)

FIGURE 11.2: The points a and b are in the same component of the complement of S, but b and c are in different components.

EXAMPLE 11.1. Consider the Jordan Curve Zoo in Figure 11.3. Is the zookeeper safe from the lion? Drawing a horizontal line L from the lion to the zookeeper, we see that L crosses the cage five times, and therefore the lion and the zookeeper are on opposite sides of the cage. If we consider the line L' running from the zookeeper to a point that we know is outside of the cage, we see that L' intersects the cage three times, and therefore we can conclude that the zookeeper is actually inside the cage.
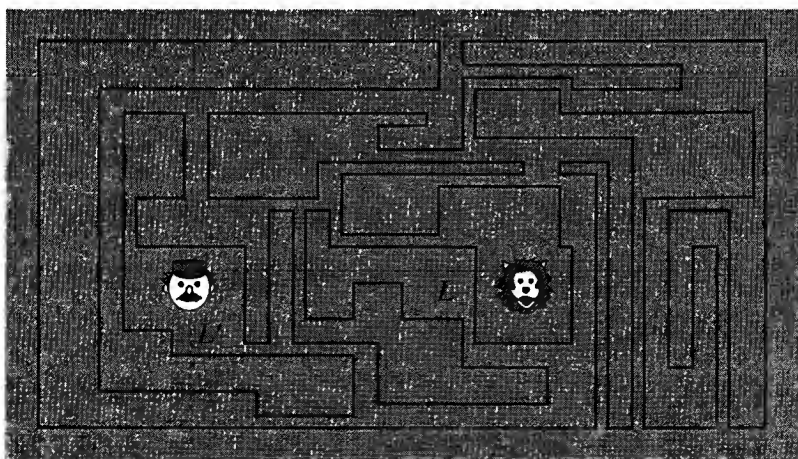
FIGURE 11.3: Is the zookeeper safe from the lion?

For simple closed curves in the plane, we can do better than the Jordan Curve Theorem. The following theorem indicates that not only is a simple closed curve homeomorphic to S^1, but the entire topological structure of a simple closed curve sitting in the plane is topologically equivalent to the circle S^1 sitting in the plane.

THEOREM 11.4. *The Schönflies Theorem.* *Let $f : S^1 \to \mathbb{R}^2$ be an embedding. Then there exists a homeomorphism $F : \mathbb{R}^2 \to \mathbb{R}^2$ such that $F|_{S^1} = f$.*

This theorem is named after Arthur Schönflies (1853–1928), who first presented it in 1906. We do not prove it; a proof can be found in [Moi]. The Schönflies Theorem states that every embedding of the circle into the plane extends to a homeomorphism of the whole plane to itself. (See Figure 11.4.) Such a homeomorphism must map the inside of S^1 in the plane to the inside of the simple closed curve $f(S^1)$ and the outside of S^1 to the outside of $f(S^1)$. (See Exercise 11.6.) Consequently, given a simple closed curve S in the plane, the inside of S is homeomorphic to the open ball $B = \{(x, y) \mid x^2 + y^2 < 1\}$ as a subspace of the plane, and the outside of S is homeomorphic to the subspace $C = \{(x, y) \mid x^2 + y^2 > 1\}$. Thus the Schönflies Theorem strengthens the Jordan Curve Theorem, indicating exactly (topologically) what the two components are that result from the simple closed curve separating the plane.
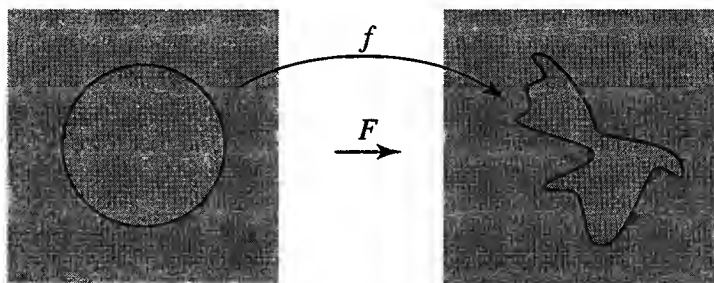
FIGURE 11.4: The embedding $f : S^1 \to \mathbb{R}^2$ extends to a homeomorphism $F : \mathbb{R}^2 \to \mathbb{R}^2$.

If we increase the dimension of the embedding space by one, and look at simple closed curves in $\mathbb{R}^3$, we find a very different situation. As indicated in Section 4.2, simple closed curves in $\mathbb{R}^3$ are known as knots. The topological possibilities for knots are endless and are the subject of a very active area of study in topology. We explore knots further in Chapter 12.

We can see from considerations of embeddings of S^1 into $\mathbb{R}^2$ and into $\mathbb{R}^3$ that mathematics often presents a win-win situation. On the one hand we have a significant theorem (the Schönflies Theorem) indicating that all embeddings of S^1 into $\mathbb{R}^2$ are essentially the same. On the other hand, it is not the case that all embeddings of S^1 into $\mathbb{R}^3$ are the same, and therefore the task of determining all of the possibilities for these embeddings develops into an important area of mathematical research.

Next, let us consider some of our familiar surfaces: the torus, sphere, Klein bottle, projective plane, and Möbius band. The sphere and torus were both initially defined as subspaces of 3-space, so they trivially embed in 3-space.

The usual picture of the Möbius band illustrates an embedding in $\mathbb{R}^3$. In fact we can define a specific embedding of the Möbius band. We view the Möbius band as the rectangle $R = [0, 2\pi] \times [-1, 1]$ with the pairs of points of

the form $(0, a)$ and $(2\pi, -a)$ identified. Using (r, θ, z) cylindrical coordinates in $\mathbb{R}^3$, we define $f : R \to \mathbb{R}^3$ by $f(s, t) = (r(s, t), \theta(s, t), z(s, t))$, where

$$r(s, t) = 5 + t\cos(s/2), \quad \theta(s, t) = s, \quad z(s, t) = t\sin(s/2).$$

Since $f(0, a) = f(2\pi, -a)$, it follows that f defines a function f^* mapping the Möbius band to $\mathbb{R}^3$. The function f^* is an embedding. It takes the core circle in the Möbius band, where $t = 0$, to the circle of radius five in the xy-plane. For each fixed $s \in [0, 2\pi]$, f^* maps the line segment $\{s\} \times [-1, 1]$ to a line segment in the plane $\theta = s$. As s runs from 0 to 2π, this segment essentially rotates halfway around, sweeping out the image of the Möbius band. (See Exercise 11.8.)

Consider the Klein bottle and the projective plane. As we indicated previously, neither can be embedded in $\mathbb{R}^3$. However, with just one more dimension at our disposal (that is, if we consider $\mathbb{R}^4$), we can define embeddings. We show how for the Klein bottle; a similar approach can be used for the projective plane. (See Exercise 11.9.)

The usual representation of the Klein bottle K in $\mathbb{R}^3$ can be viewed as the image of a function $f : K \to \mathbb{R}^3$. (See Figure 11.5.) The function f is not an embedding because it is not one-to-one as a result of the self-intersection. We call this the standard $\mathbb{R}^3$ representation of K.
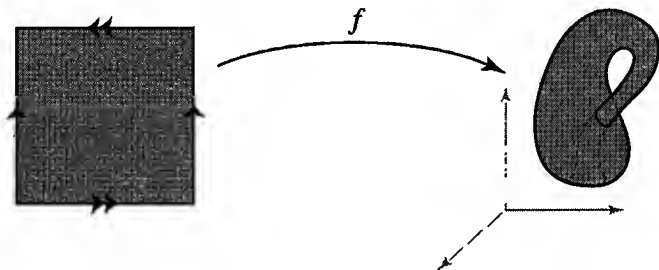
FIGURE 11.5: Mapping the Klein bottle into $\mathbb{R}^3$.

In Figure 11.6 we depict the Klein Bottle as a square with its edges identified. Within that depiction we have illustrated the set D that consists of the points on the Klein bottle where the self-intersection occurs in the standard $\mathbb{R}^3$ representation (that is, the points in K where f fails to be one-to-one). We define a continuous function $h : K \to \mathbb{R}$ by choosing h to be constant on horizontal lines in the square, and by having h increase from 0 to 1 and then decrease back to 0 as we move vertically from the bottom to the top of the square. The idea is to have $h(x_1) \neq h(x_2)$ if x_1 and x_2 are points in D corresponding to the same point in the standard $\mathbb{R}^3$ representation of K. The function h serves as one coordinate of the embedding of K into $\mathbb{R}^4$.

Define a function $F : K \to \mathbb{R}^4$ by having the first three coordinates of $F(x)$ in $\mathbb{R}^4$ equal $f(x)$ and the last coordinate equal $h(x)$. In this case, F is one-to-one because if x and y are such that $x \neq y$ and $f(x) = f(y)$, then the
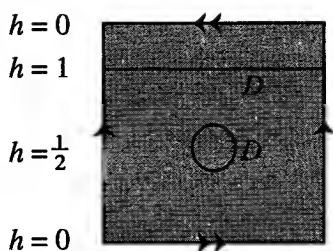
FIGURE 11.6: Defining $h : K \to \mathbb{R}$, one coordinate of an embedding of K into $\mathbb{R}^4$.

function h is defined so that $h(x) \neq h(y)$. Therefore for every $x, y \in K$, if $x \neq y$ then $F(x) \neq F(y)$, and it follows that the function F is an embedding of K in $\mathbb{R}^4$.

It is natural to ask if there are higher-dimensional analogs of the Jordan Curve Theorem and the Schönflies Theorem. For example, does every embedded sphere in $\mathbb{R}^3$ separate $\mathbb{R}^3$ into two components with the sphere being the boundary of each (as in the Jordan Curve Theorem)? And, if so, are the two components topologically equivalent to the two components resulting from the standard embedding (as in the Schönflies Theorem)? The answer to the first of these questions is yes, but to the second is no.

First, let us address the affirmative answer. Not only does the Jordan Curve Theorem extend to embeddings of the sphere in $\mathbb{R}^3$, but it also generalizes to all dimensions:

THEOREM 11.5. *The Jordan–Brouwer Separation Theorem.* *Let S be an embedded image of S^n in $\mathbb{R}^{n+1}$. Then $\mathbb{R}^{n+1} - S$ consists of two components, and S is the boundary in $\mathbb{R}^{n+1}$ of each of the components.*

L. E. J. Brouwer proved this result in 1912. This theorem is straightforward to prove, given appropriate tools from algebraic topology. (See [Vic], for example.) A proof that does not require tools from algebraic topology can be found in [Dug].

The question of whether or not it is possible to generalize the Schönflies Theorem to higher dimensions was not resolved until 1924, when J. W. Alexander (1888–1971) discovered a counterexample, now known as the Alexander Horned Sphere. The example consists of a sphere embedded in $\mathbb{R}^3$ such that the outside of the sphere is not homeomorphic to the outside of S^2 in $\mathbb{R}^3$.

The Alexander Horned Sphere is the embedded sphere A shown in Figure 11.7. Between the two main horns seen in the figure, there is intricate detail that we describe and illustrate further in a moment. The loop L in Figure 11.7 lies in the outside of A, but it cannot be deformed to a point in the outside of A (we discuss why shortly). Therefore the outside of A is not simply connected. On the other hand, the outside of S^2 is homeomorphic to $\mathbb{R}^3 - \{O\}$, which is simply connected by Theorem 9.20. Thus, the embedding of S^2 to the Alexander Horned Sphere cannot be extended to a homeomorphism of $\mathbb{R}^3$ to itself. It follows that the Schönflies Theorem does not hold in dimension three.
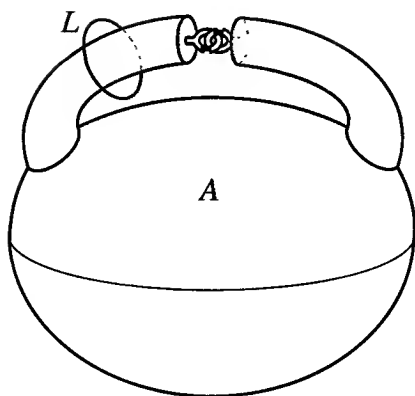
FIGURE 11.7: The Alexander Horned Sphere.

To construct the Alexander Horned Sphere, we begin with a sphere and push out two horns as shown on the left in Figure 11.8. Next, in the second step in the construction, we push out a pair of hooked horns in the gap between the ends of the first two pushed-out horns, as shown on the right in the figure. The result is still an embedded sphere because there is a gap between the ends of each horn and they do not close up.
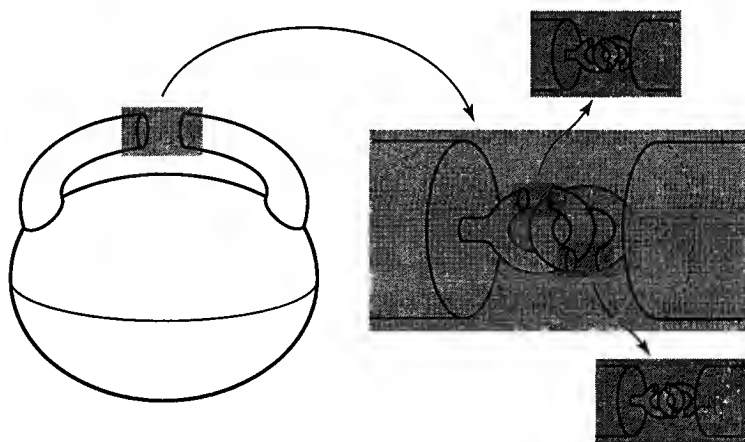
FIGURE 11.8: Pushing out the Alexander Horned Sphere horns.

We continue this construction process; as shown on the right in Figure 11.8, in each step we push out two smaller hooked horns in the gap between the ends of each pair of horns formed in the previous step. At the end of the nth step we have 2^{n-1} gaps, and therefore in the next step we push out 2^n more hooked horns. The result, after carrying out this process for all $n \in \mathbb{Z}_+$, is the Alexander Horned Sphere.

Now, why is it that the loop L in Figure 11.7 cannot be deformed to a point in the outside of A? Roughly, the idea is as follows: If we could deform L to a point, then we would have a continuous function F mapping the disk D into the

outside of A that, when restricted to its boundary, S^1, is an embedding of the circle onto L. Such a function would need to have the image of the disk pass between the hooked horns in A. Since D is compact, $F(D)$ is a compact subset of $\mathbb{R}^3$. Also, $F(D)$ is disjoint from A, which is a compact set in $\mathbb{R}^3$. These two disjoint compact sets in $\mathbb{R}^3$ must have a positive distance, say ε, between them by Theorem 7.25. In the construction of A, the hooks eventually get to the point where the gaps between them are smaller than ε, and therefore the image of F cannot get through those gaps to pass between the hooked horns.

Exercises for Section 11.1

11.1. (a) Prove that there is no embedding $f : S^1 \to \mathbb{R}$.
   (b) Prove that an embedding $f : S^1 \to S^1$ must be surjective and therefore must be a homeomorphism.

11.2. Prove that being embeddable into a particular space is a topological invariant. Specifically, prove that if Y is a topological space and X and X' are homeomorphic, then X embeds in Y if and only if X' embeds in Y.

11.3. (a) Use the Schönflies Theorem to prove that every simple closed curve in the sphere separates the sphere into two components, each of which is homeomorphic to an open disk.
   (b) On each of the torus, the Klein bottle, and the projective plane, sketch two simple closed curves, one that separates the space and one that does not.

11.4. Prove the Schönflies Theorem for embeddings $f : S^0 \to \mathbb{R}$. That is, prove that if $f : S^0 \to \mathbb{R}$ is an embedding, then f extends to a homeomorphism $F : \mathbb{R} \to \mathbb{R}$.

11.5. An embedding $f : S^1 \to \mathbb{R}^2$ is called a **star embedding** if there exists a point p on the inside of $f(S^1)$ and a continuous function $r : S^1 \to (0, \infty)$ such that, with polar coordinates centered at p, f can be expressed as $f(\theta) = (r(\theta), \theta)$. (See Figure 11.9.) Prove the Schönflies Theorem for star embeddings.
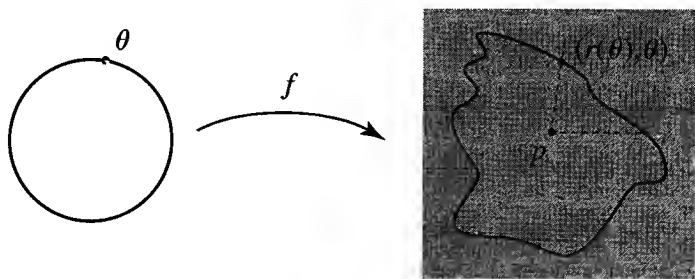
FIGURE 11.9: A star embedding.

11.6. Show that if $f : \mathbb{R}^2 \to \mathbb{R}^2$ is a homeomorphism, then f maps the inside of S^1 to the inside of the simple closed curve $f(S^1)$ and maps the outside of S^1 to the outside of $f(S^1)$.

11.7. Show that the Schönflies Theorem implies that if $f : D \to \mathbb{R}^2$ is an embedding, then $\mathbb{R}^2 - f(D)$ is connected. (Therefore, no subspace of the plane homeomorphic to a disk separates the plane.)

11.8. Illustrate the embedding of the Möbius band in $\mathbb{R}^3$ given by the function f^* described in this section. In particular, illustrate how the image of f^* appears on the cylindrical-coordinate planes of fixed angle θ, as θ runs from 0 to 2π.

11.9. Describe how the projective plane can be embedded in $\mathbb{R}^4$.

11.2   *The Jordan Curve Theorem*

In this section we prove the Jordan Curve Theorem. This intuitively straightforward result is surprisingly difficult to prove. Many different proofs, with a variety of different approaches, have been published over the past century since the theorem was first proved by Veblen in 1905. The Jordan Curve Theorem may be the most-proved theorem in topology. The approach that we present uses the two-dimensional versions of the No Retraction Theorem and the Brouwer Fixed Point Theorem (which, as shown in Section 10.1, are equivalent theorems). It also uses Theorem 9.14, which asserts the existence of retractions from the disk onto arc subspaces. Our proof is based on the proof in [Mae].

Along with the proof of the Jordan Curve Theorem, we also present a nonseparation theorem (Theorem 11.10) indicating that no arc in the plane separates the plane. This result is essentially proven in the process of establishing the Jordan Curve Theorem.

Before we prove the Jordan Curve Theorem, we present a few results that we need in the proof.

We begin with a theorem that states that if we have a rectangle with a path running from the left side to the right side and a path running from the bottom to the top, then the paths must have a point in common. The Two-Dimensional Brouwer Fixed-Point Theorem is used to prove this result.

Previously, we defined a path in a topological space X to be a continuous function $f : [0, 1] \to X$. Here, for convenience, we regard paths as having domain $[-1, 1]$, rather than $[0, 1]$. This does not pose a problem since the spaces $[0, 1]$ and $[-1, 1]$ are homeomorphic. As before, we also use the term path to refer to the image of the function that is the path.

Consider the rectangle $R = [-1, 1] \times [-1, 1]$. Let B be the boundary of R as a subset of the plane; so $B = \{(s, t) \in R \mid |s| = 1 \text{ or } |t| = 1\}$.

THEOREM 11.6. *Let $u, v : [-1, 1] \to R$ be paths in R such that $u(-1) = (-1, 0)$, $u(1) = (1, 0)$, $v(0) = (0, -1)$, and $v(1) = (0, 1)$. Then there exist $s, t \in [-1, 1]$ such that $u(s) = v(t)$. (See Figure 11.10.)*

> *Proof.* Assume that no such s and t exist, and therefore $u(s) \neq v(t)$ for all $s, t \in [-1, 1]$. Expressing $u(s)$ and $v(t)$ in terms of their coordinates, we have $u(s) = (u_1(s), u_2(s))$ and $v(t) = (v_1(t), v_2(t))$. For each $s, t \in [-1, 1]$, we define
>
> $$M(s, t) = \max\{|u_1(s) - v_1(t)|, |u_2(s) - v_2(t)|\}.$$
>
> Note that $M(s, t) > 0$ for all s and t.
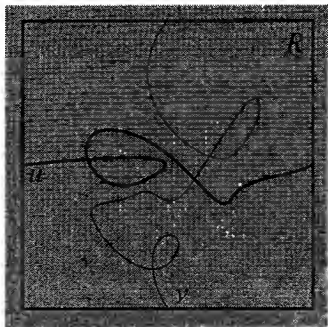
FIGURE 11.10: Paths u and v in R must cross.

Next, define $F : R \to R$ by

$$F(s, t) = \left(\frac{v_1(t) - u_1(s)}{M(s, t)}, \frac{u_2(s) - v_2(t)}{M(s, t)} \right).$$

The function F is continuous, and R is homeomorphic to the disk; therefore it follows by the Brouwer Fixed Point Theorem that F must have a fixed point $(s^*, t^*) \in R$. Note that for each $s, t \in [-1, 1]$, either

$$\left| \frac{v_1(t) - u_1(s)}{M(s, t)} \right| = 1 \quad \text{or} \quad \left| \frac{u_2(s) - v_2(t)}{M(s, t)} \right| = 1.$$

Therefore the image of F is a subset of B, the boundary of the rectangle R. Thus, the fixed point (s^*, t^*) must lie in B. We have four possibilities to consider: $s^* = -1$, $s^* = 1$, $t^* = -1$, or $t^* = 1$. We claim that each of the four cases leads to a contradiction, and therefore there must be $s, t \in [-1, 1]$ such that $u(s) = v(t)$. Here we derive a contradiction in the case that $s^* = -1$. The remaining cases are addressed in Exercise 11.12.

Thus assume $s^* = -1$. Then $\frac{v_1(t^*) - u_1(s^*)}{M(s^*, t^*)} = -1$. But

$$(u_1(s^*), u_2(s^*)) = u(s^*) = u(-1) = (-1, 0),$$

and therefore $u_1(s^*) = -1$. This implies that $\frac{v_1(t^*) - u_1(s^*)}{M(s^*, t^*)} \geq 0$, which is a contradiction.   ∎

THEOREM 11.7. *Let U be an open subset of the plane. Then U is connected if and only if it is path connected.*

Proof. By Theorem 6.28, path connectedness implies connectedness; consequently we only need to prove that if U is connected, then it is path connected. Therefore, suppose that U is connected. If U is empty, then the result trivially holds; thus assume that U is nonempty, and pick a point $p \in U$. Let Y_p be the set of points q in U such that there exists a path in U from p to q, and let N_p be the set of points q in U such that no path exists in U from p to q. The sets Y_p and N_p are disjoint and their union is U.

We claim that Y_p and N_p are both open sets. To prove the claim, we prove that Y_p is open; the proof for N_p is similar. Suppose $q \in Y_p$. Then there is a path f in U from p to q. Since U is open, it follows that in the standard metric on $\mathbb{R}^2$ there is an open ball B that is centered at q and contained in U. Pick a point $q' \in B$. There is a path g in B (and therefore in U) from q to q'. Define a path h by having it follow f from p to q and then follow g from q to q'. The path h is a path in U from p to q'. Therefore every q' in B is also in Y_p. Thus, for arbitrary $q \in Y_p$ there is an open ball B such that $q \in B \subset Y_p$. It follows that Y_p is an open set.

Now, Y_p and N_p are disjoint open sets whose union is U. If Y_p and N_p were both nonempty, then they would form a separation of U, which is not possible since U is connected. Therefore either Y_p or N_p is empty. Since $p \in Y_p$, it must be N_p that is empty. This implies that $U = Y_p$, and thus there exists a path in U from p to every other point in U. Hence, U is path connected. ∎

THEOREM 11.8. *Let U be an open subset of the plane. Then the components of U are open subsets of the plane.*

Proof. See Exercise 11.10. ∎

By combining Theorems 11.7 and 11.8, it follows that if U is an open subset of the plane, then the components and path components of U coincide and are open subsets of the plane.

The last theorem that we need before beginning our proof of the Jordan Curve Theorem is the following consequence of the Two-Dimensional No Retraction Theorem:

THEOREM 11.9. *Let D be the disk in the plane. If $f : D \to D$ is continuous and $f(x) = x$ for all $x \in S^1$, then f is surjective.*

Proof. See Exercise 11.11. ∎

Theorem 11.9 indicates that if f is a continuous function that maps the disk to itself and is the identity on its boundary—the circle S^1—then f must map onto the disk.

Now we are ready to prove the Jordan Curve Theorem, which states that if S is a simple closed curve in the plane, then $\mathbb{R}^2 - S$ has two components and S is the boundary of each component.

We prove this in three parts: first we show that $\mathbb{R}^2 - S$ is not connected, then we show that S is the boundary of each component of $\mathbb{R}^2 - S$, and finally we show that $\mathbb{R}^2 - S$ has two components. A few straightforward verifications in the proofs are addressed in the exercises.

Proof ($\mathbb{R}^2 - S$ *is not connected*). Let $d(x, y) : S \times S \to \mathbb{R}$ be the function defining the distance between x and y in the standard metric on the plane. The domain of d is compact and d is continuous; therefore d takes

on a maximum value on $S \times S$. It follows that there is at least one pair of points s_1, s_2 in S that are a maximum distance apart. Let J be the line segment in the plane connecting s_1 and s_2, and let P_1 and P_2 be the lines perpendicular to J passing through s_1 and s_2, respectively. (See Figure 11.11.) The points s_1 and s_2 are the only points of S lying on $P_1 \cup P_2$, and the rest of S lies between P_1 and P_2. (See Exercise 11.13.) Since S is compact and therefore bounded, there exist lines J_1 and J_2, parallel to and equidistant from J, such that $S - \{s_1, s_2\}$ lies in the interior of the bounded rectangle R' bordered by P_1, P_2, L_1, and L_2.
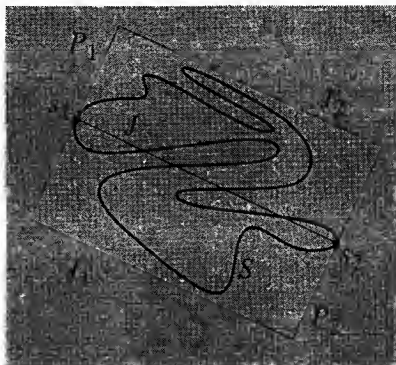
FIGURE 11.11: The simple closed curve S lies in the rectangle bordered by P_1, P_2, L_1, and L_2.

By rotating, translating, and scaling we can obtain a homeomorphism h from the plane to itself, mapping R' to $R = [-1, 1] \times [-1, 1]$ and sending s_1 and s_2 to $(-1, 0)$ and $(1, 0)$, respectively. (See Figure 11.12.) Thus, it suffices to assume that S lies in the rectangle R and that S intersects the boundary of R only at the points $s_1 = (-1, 0)$ and $s_2 = (1, 0)$.
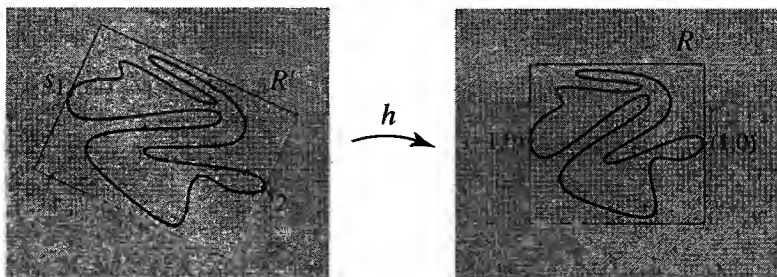
FIGURE 11.12: The rectangle R' maps homeomorphically to $[-1, 1] \times [-1, 1]$.

Now, S forms two arcs in R, running from $s_1 = (-1, 0)$ to $s_2 = (1, 0)$. We denote these arcs by S' and S''. Consider the vertical line segment L in R connecting $a = (0, -1)$ to $b = (0, 1)$. Using Theorem 11.6, we can conclude that L intersects both S' and S''. Traveling from a to b

along L, we may assume, without loss of generality, that L intersects S' before it intersects S''. Let a' be the first point of intersection of L with S' as we go from a to b along L, and let b' be the last point of intersection. (See Figure 11.13.)
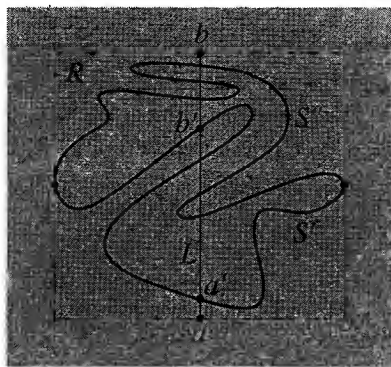
FIGURE 11.13: Going from a to b along L, the points a' and b' are the first and last intersections, respectively, with S'.

We claim that L intersects S'' between b' and b. If it did not, then we could take a path, denoted $aLa'S'b'Lb$, that goes from a to a' along L, then from a' to b' along S', and finally from b' to b along L. Such a path would go from a to b in R without intersecting S''. (See Exercise 11.14.) This is impossible by Theorem 11.6. Therefore L intersects S'' between b' and b. Now, as we go from b' to b along L, let a'' be the first point of intersection between L and S'' and let b'' be the last.

Consider the point c lying halfway between b' and a'' on L. It follows that c lies on neither S' nor S''. Therefore c is in the complement of S in $\mathbb{R}^2$. Let U be the component of $\mathbb{R}^2 - S$ that contains c; we claim that U is bounded.

To show that U is bounded, suppose that it is not. By Theorems 11.7 and 11.8, U is path connected, and since U is unbounded, we can find a path P in U from c to the outside of R. Let p' be the first point on P where P intersects the boundary of R. The point p' is in either the bottom half or the top half of the boundary of R. Suppose that p' is in the bottom half. Let Q be a path in the boundary of R running from a to p'. Using the same notational convention as above, a path $aQp'PcLb$ goes from the bottom to the top of R and does not intersect S'. (See Exercise 11.14.) This contradicts Theorem 11.6. Next suppose that p' is in the top half of the boundary of R. Let Q' be a path in the boundary of R running from p' to b. Then a path $aLa'S'b'LcPp'Q'b$ goes from the bottom to the top of R and does not intersect S''. (See Exercise 11.14.) This also yields a contradiction to Theorem 11.6. Thus, in either case—p' in the bottom half of the boundary of R, or p' in the top half of the boundary of R—we have derived a contradiction, and therefore U is bounded.

Since S is a bounded subset of the plane, it follows that $\mathbb{R}^2 - S$ has an unbounded component. Therefore $\mathbb{R}^2 - S$ has at least two components, one bounded and one unbounded, implying that $\mathbb{R}^2 - S$ is not connected.

∎

In the next part of the proof of the Jordan Curve Theorem, we show that S is the boundary of each component of $\mathbb{R}^2 - S$.

Proof (S is the boundary of each component of $\mathbb{R}^2 - S$).   To   begin, note that Theorem 11.8 implies that each component of $\mathbb{R}^2 - S$ is an open set. Let V be such a component. Since the other components of $\mathbb{R}^2 - S$ are all open sets, it follows that $V \cup S$ is closed and therefore $\mathrm{Cl}(V) \subset V \cup S$. Now, $\mathrm{Int}(V) = V$ since V is open, and therefore $\mathrm{Cl}(V) - \mathrm{Int}(V) \subset S$, implying that $\partial V \subset S$. Therefore, the boundary of each component of $\mathbb{R}^2 - S$ is a subset of S. We need to show that the boundary of each component of $\mathbb{R}^2 - S$ equals S. We do so by contradiction.

Thus let V be a component of $\mathbb{R}^2 - S$ and assume that the boundary of V does not equal S. Therefore ∂V is a closed proper subset of S. It follows that there is an arc A in S containing ∂V. We consider the cases where V is bounded and where V is unbounded separately.

First assume that V is bounded. Let B be a closed ball in the plane that is large enough to contain $V \cup S$ in its interior. By Theorem 9.14, there is a retraction $r : B \to A$. Define $f : B \to B$ by

$$f(x) = \begin{cases} r(x) & \text{if } x \in \mathrm{Cl}(V), \\ x & \text{if } x \in B - V. \end{cases}$$

Note that $\mathrm{Cl}(V)$ and $B - V$ are closed subsets of B. Furthermore, $\mathrm{Cl}(V) \cap (B - V) \subset A$ and $r(x) = x$ for $x \in A$. The Pasting Lemma implies that f is continuous. Now, f equals the identity when restricted to the boundary circle of B. Therefore Theorem 11.9 implies that f is surjective. However, it follows from the definition of f that no point in V is in the image of f, and thus we have a contradiction.

Now assume that V is not bounded. Let B be a closed ball in the plane that is large enough to contain in its interior the simple closed curve S and all of the bounded components of $\mathbb{R}^2 - S$. By the first part of the proof, there is at least one bounded component of $\mathbb{R}^2 - S$; denote it by U. Again, by Theorem 9.14, there is a retraction $r : B \to A$. Define $g : B \to B$ by

$$g(x) = \begin{cases} x & \text{if } x \in \mathrm{Cl}(V) \cap B, \\ r(x) & \text{if } x \in B - V. \end{cases}$$

The function g is continuous and equals the identity when restricted to the boundary circle of B. Theorem 11.9 implies that g is surjective, but it follows from the definition of g that no point in U is in the image of g. Therefore, we have a contradiction in this case as well.

Thus S is the boundary of each component of $\mathbb{R}^2 - S$.   ∎

In the last part of the proof of the Jordan Curve Theorem, we show that $\mathbb{R}^2 - S$ has two components. We use the same setup and notations as in the first part of the proof.

> ***Proof*** *($\mathbb{R}^2 - S$ **has two components**).* Let U be the bounded component of $\mathbb{R}^2 - S$ whose existence was established in the first part of the proof. Assume that there is another bounded component, W. We derive a contradiction that allows us to conclude that $\mathbb{R}^2 - S$ has one bounded component. Since $\mathbb{R}^2 - S$ also has one unbounded component, it then follows that $\mathbb{R}^2 - S$ has two components.
>
> Consider the arc A in R given by $aLa'S'b'La''S''b''Lb$. It runs from the bottom of R to the top and is disjoint from W. (See Exercise 11.15.) The points $s_1 = (-1, 0)$ and $s_2 = (1, 0)$ do not lie in A. Since A is closed, there exist open balls B_1 and B_2 in the plane, centered at s_1 and s_2, respectively, that are disjoint from A. By the second part of the proof, the points s_1 and s_2 lie in the boundary of W. Therefore there exist points w_1 and w_2 in $B_1 \cap W$ and $B_2 \cap W$, respectively. Let M be a path in B_1 from s_1 to w_1, let M' be a path in W from w_1 to w_2, and let M'' be a path in B_2 from w_2 to s_2. Then a path $s_1 M w_1 M' w_2 M'' s_2$ is a path in R that goes from s_1 to s_2 and does not intersect A. This contradicts Theorem 11.6. Thus there is only one bounded component in $\mathbb{R}^2 - S$, and the proof of the Jordan Curve Theorem is complete. ∎

The following is a nonseparation theorem that is closely related to the Jordan Curve Theorem. We ask you to prove it in Exercise 11.16.

THEOREM 11.10. *Let A be an arc in the plane. Then $\mathbb{R}^2 - A$ is connected, and A is the boundary of $\mathbb{R}^2 - A$.*

> ***Proof.*** See Exercise 11.16. ∎

Exercises for Section 11.2

11.10. Prove Theorem 11.8: Let U be an open subset of the plane. Then the components of U are open subsets of the plane.

11.11. Prove Theorem 11.9: Let D be the disk in the plane. If $f : D \to D$ is continuous and $f(x) = x$ for all $x \in S^1$, then f is surjective. (Hint: Use the No Retraction Theorem.)

11.12. In the proof of Theorem 11.6, show that you arrive at a contradiction in each of the cases $s^* = 1$, $t^* = -1$, and $t^* = 1$, for the coordinates of the fixed point (s^*, t^*) of the function F.

11.13. For the Jordan Curve Theorem proof that $\mathbb{R}^2 - S$ is not connected, provide arguments for the claims that the points s_1 and s_2 are the only points of S lying on $P_1 \cup P_2$ and that the rest of S lies between P_1 and P_2.

11.14. For the Jordan Curve Theorem proof that $\mathbb{R}^2 - S$ is not connected, provide arguments for the following claims:
  (a) A path $aLa'S'b'Lb$ does not intersect S''.
  (b) A path $aQp'PcLb$ does not intersect S'.
  (c) A path $aLa'S'b'Lc Pp'Q'b$ does not intersect S''.

11.15. For the Jordan Curve Theorem proof that $\mathbb{R}^2 - S$ consists of two components, provide the argument for the claim that the arc $a\,La'\,S'b'\,La''\,S''b''\,Lb$ is disjoint from W.

11.16. Prove Theorem 11.10: Let A be an arc in the plane. Then $\mathbb{R}^2 - A$ is connected, and A is the boundary of $\mathbb{R}^2 - A$. This is done in three parts:
  (a) Prove that if C is a closed subset of a topological space X, then the boundary of the complement of C is a subset of C. Use this to show that $\partial(\mathbb{R}^2 - A) \subset A$.
  (b) Prove that $\mathbb{R}^2 - A$ is connected. (Hint: Use an argument similar to one used in the Jordan Curve Theorem proof that S is the boundary of each component of $\mathbb{R}^2 - S$.)
  (c) Prove that $A \subset \partial(\mathbb{R}^2 - A)$. (Hint: Assume that A is not a subset of $\partial(\mathbb{R}^2 - A)$, and argue why there must be a point in A with an open–ball neighborhood that is disjoint from $\mathbb{R}^2 - A$. Then, explain why that is an impossibility.)

11.3 *Digital Topology and Digital Image Processing*

Digital images have become one of the primary means for communicating visual information. The pictures in a digital camera, the illustrations in this book, the graph on a calculator screen, and the display on a ballpark scoreboard are examples of images that are digitally constructed or presented. The field of digital image processing is concerned with the creation, storage, manipulation, and presentation of digital images. In each of these aspects of digital image processing, there are questions and problems that involve topological concepts and tools. For example,

  (i) In creating a digital image, how can it be ensured that spatial relationships between real-world features are properly depicted in a digital representation of them?

  (ii) In storing a digital image, are there characteristics of the image's structure that allow for a more efficient method than storing information about every individual pixel?

  (iii) In digitally transforming an image, how can the topological aspects of the image be preserved?

In this section we consider question (ii) and present topological results that suggest an efficient storage process for digital images. These results include a digital version of the Jordan Curve Theorem. Our presentation is based on results from [Kha3].

The topological setting for this section is the digital plane, a space obtained by taking the product of two digital lines. We introduced the digital plane in Section 1.4, where we presented it as the topological space defined on $\mathbb{Z} \times \mathbb{Z}$ via the basis of sets

$$B(m, n) = \begin{cases} \{(m, n)\} & \text{if } m \text{ and } n \text{ are odd,} \\ \{(m + a, n) \mid a = -1, 0, 1\} & \text{if } m \text{ is even and } n \text{ is odd,} \\ \{(m, n + b) \mid b = -1, 0, 1\} & \text{if } m \text{ is odd and } n \text{ is even,} \\ \{(m + a, n + b) \mid a, b = -1, 0, 1\} & \text{if } m \text{ and } n \text{ are even.} \end{cases}$$

We denote the digital plane by $\mathbb{Z}^2$. For $(m, n) \in \mathbb{Z}^2$, we refer to $B(m, n)$ as the **minimal basis element containing** (m, n). We illustrate some of these basis elements in Figure 11.14. In the illustration we use different symbols for each of the different types of points in the digital plane. Points of the form (m, n) with both m and n odd are called **open points**. We denote them by O. Each of these points is a single-point open set in the digital plane. Points (m, n), with both m and n even, are called **closed points**. We denote the closed points by ✳, reflecting the fact that the minimal basis element for a point ✳ contains the eight points surrounding it. Points (m, n), with one odd and one even coordinate, are called **mixed points**. We denote the mixed points by either ◆ or ◗, with ◆ reflecting the fact that the minimal basis element for a point ◆ contains the points directly above and below it. A similar situation holds for the points denoted ◗.
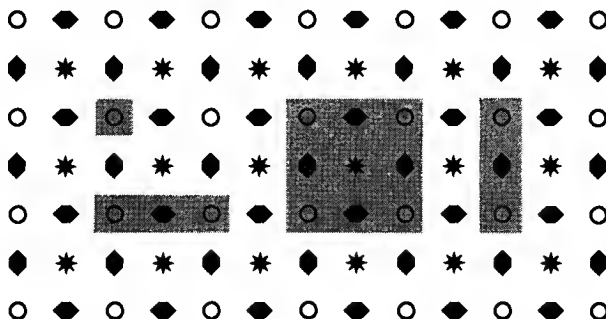
FIGURE 11.14: Basis elements in the digital plane.

The digital plane can also be obtained as a quotient space of the plane $\mathbb{R}^2$ with the standard topology. (See Exercise 3.32.) Therefore, since the plane is path connected and the digital plane is the continuous image of the plane under the corresponding quotient map, it follows that the digital plane is path connected. In fact, given two points in the digital plane, we can define a path $f : [0, 1] \to \mathbb{Z}^2$, as illustrated in Figure 11.15, that traces a route traveling vertically, horizontally, or both, from one point to the other. (See Exercise 11.17.)

The subspace of the digital plane consisting of all of the open points is called the **visible screen**; we denote it by $\mathbb{V}$. In our digital-image-processing model, the visible screen corresponds to the set of pixels in a digital image display. That is, it corresponds to what we actually see in a digital picture. The visible screen is a dense and open subset of the digital plane, and it inherits the discrete topology as a subspace. The inclusion of the closed points and the mixed points provides an invisible structure that connects the pixels and allows for the use of topological concepts and results in modeling and studying properties of digital images.
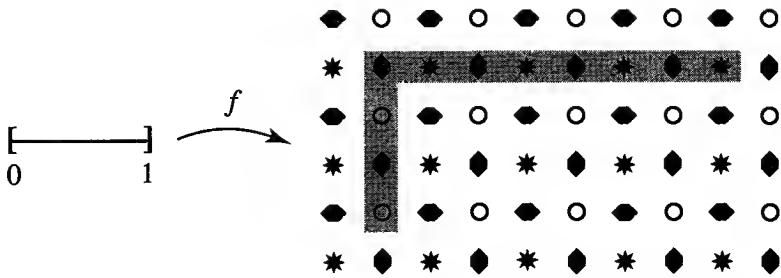
FIGURE 11.15: Given two points in the digital plane, there is a path between them.

With the digital plane as the setting for our model, let us consider the storage problem previously introduced. In Figure 11.16, shaded in gray, we present sets of digital-plane points that enclose 1×1, 2×2, and 3×3 arrays of pixels in the visible screen.
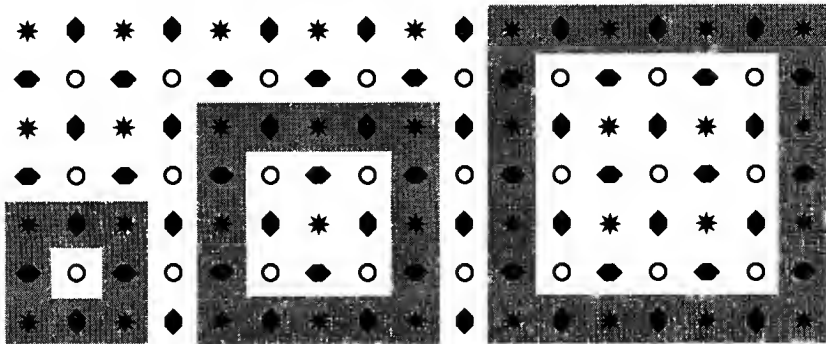
FIGURE 11.16: Digital-plane sets enclosing pixel arrays.

The surrounding sets in Figure 11.16 are made up of 8, 16, and 24 points, respectively. In general, it can be shown that an $n \times n$ array of open points can be enclosed by a surrounding set of $8n$ points in the digital plane. (See Exercise 11.23.) Therefore, if we want to store a digital image of a square blue area, 1000×1000 pixels in size, we can either store the location of each of the one-million blue-colored pixels, or we can store the location of the 8000 surrounding points in the digital plane, indicating that the corresponding enclosed points in the visible screen are colored blue. Clearly, there is a storage savings in working with the surrounding set rather than the entire enclosed set.

Now, to what extent does this principle carry over to more general digital images? We model a digital picture as a collection of regions such that in each region every point has the same color. We would like to know if we can replace the regions with surrounding sets that uniquely determine the regions. In this way we could reduce the space needed to store a picture by storing only information about the surrounding sets and the colors. Since the picture's regions are uniquely determined by the surrounding sets, we could recover the picture from the stored information. The remainder of this section is an introduction to some digital-topology results that help answer this question.

The surrounding sets in Figure 11.16 each inherit a topology as a subspace of the digital plane. It is easy to see that each of these subspaces is a digital circle, as introduced in Section 3.3. Digital circles C_n are defined for even integers $n \geq 4$. They are obtained by taking the subspace $\{1, ..., n + 1\}$ of the digital line and gluing the ends 1 and $n + 1$ together. In Figure 11.17 we show the digital circles C_4, C_6, and C_8, along with a basis for each.
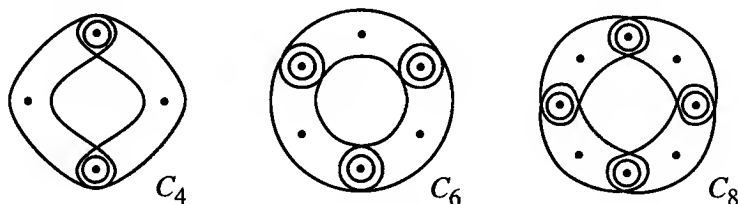
FIGURE 11 17: Digital circles C_4, C_6, and C_8.

DEFINITION 11.11. *Let X be a topological space. A **digital simple closed curve** in X is a subspace of X that is homeomorphic to a digital circle.*

EXAMPLE 11.2. The four-point set A in Fig. 11.18 is a digital simple closed curve, but the four-point set B is not. The subspace topology that B inherits from the digital plane is the discrete topology.
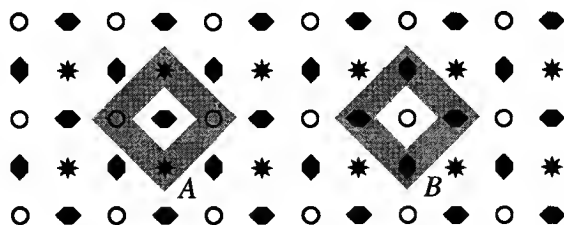
FIGURE 11.18: The set A is a digital simple closed curve, but the set B is not.

The Jordan Curve Theorem indicates that every simple closed curve separates the plane, $\mathbb{R}^2$, into two components, each of which has the simple closed curve as its boundary. Therefore a simple closed curve S in the plane determines two regions that are surrounded by S, and these regions each determine S since S is obtained by taking the boundary of each. This is the type of result that we seek for regions and surrounding sets in the digital plane. Is there a corresponding Jordan Curve Theorem for digital simple closed curves in the digital plane? We will see shortly that the answer is yes, but with a minor difference.

Considering the three digital simple closed curves in Figure 11.16 and the single digital simple closed curve in Figure 11.18, it is straightforward to show that each of these sets separates the digital plane into two connected subsets.

Also, in the three examples in Figure 11.16, the digital simple closed curve is the boundary of each of the components into which the set separates the digital plane. In contrast, the digital simple closed curve A in Figure 11.18 is not the boundary of each component of its complement in the digital plane. (See Exercise 11.18.)

It follows that we cannot attain a digital version of the Jordan Curve Theorem that corresponds exactly with the standard version. However, we have the following theorem:

THEOREM 11.12. *The Digital Jordan Curve Theorem.* *Let A be a digital simple closed curve in the digital plane. Then A separates the digital plane into two components. Furthermore, A is the boundary of each of the components if and only if A is a closed subset of the digital plane.*

We do not present a proof of Theorem 11.12, although it is accessible with the tools and concepts presented up to this point in the text. A proof, by induction on the number of points in the digital simple closed curve, is presented in [Kis]. The initial case in the induction proof addresses four-point digital simple closed curves. That case is a consequence of the following theorem:

THEOREM 11.13. *Every four-point digital simple closed curve consists of points $(m-1, n)$, $(m+1, n)$, $(m, n-1)$, and $(m, n+1)$, two of which are open points and two of which are closed points.*

> *Proof.* Let A be a four-point digital simple closed curve. Using Theorem 1.13, it readily follows that the collection of sets
>
> $$\{B(p, q) \cap A \mid (p, q) \in A\}$$
>
> is a basis for the subspace topology on A. Since A is homeomorphic to the digital circle C_4, this basis must include two three-point sets that intersect in a pair of points. That can only happen if A contains two closed points $(m-1, n)$ and $(m+1, n)$, or $(m, n-1)$ and $(m, n+1)$. The resulting three-point sets must then intersect in a pair of open points, either the points $(m, n-1)$ and $(m, n+1)$, or the points $(m-1, n)$ and $(m+1, n)$, respectively. ∎

It follows from Theorem 11.13 that a four-point digital simple closed curve must appear either as the set A in Figure 11.18 or as a 90-degree rotation of it. Now, suppose we have a four-point digital simple closed curve

$$A = \{(m-1, n), (m+1, n), (m, n-1), (m, n+1)\}.$$

If we let $C_1 = \{(m, n)\}$ and $C_2 = \mathbb{Z}^2 - (A \cup \{(m, n)\})$, then it is straightforward to show that C_1 and C_2 are the components of the complement of A in $\mathbb{Z}^2$. This verifies Theorem 11.12 for four-point digital simple closed curves.

The primary difference between the standard and digital versions of the Jordan Curve Theorem is that a simple closed curve in the standard plane is automatically the boundary of each component of its complement, while a digital simple closed curve in the digital plane is the boundary of each component of its complement if and only if the digital simple closed curve is a closed set. It is noteworthy that a simple closed curve in the standard plane must be a closed set. (See Exercise 11.21.) Therefore, in the standard Jordan Curve Theorem it is not relevant to address simple closed curves that are not closed sets, in contrast to the Digital Jordan Curve Theorem where both possibilities for digital simple closed curves—being closed or not—are considered.

The following theorem provides a simple condition for determining whether or not a digital simple closed curve is a closed subset of the digital plane.

THEOREM 11.14. *Let A be a digital simple closed curve in the digital plane. Then A is a closed subset of the digital plane if and only if A does not contain any open points.*

> *Proof.* See Exercise 11.22.   ∎

For example, each of the digital simple closed curves in Figure 11.16 consists of only closed points and mixed points. Therefore each is a closed subset of the digital plane. However, the digital simple closed curve A in Figure 11.18 contains two open points, and therefore A is not closed in the digital plane.

Given a digital simple closed curve in the digital plane, its complement consists of two components. We next show that, as with the standard Jordan Curve Theorem, one of these components is bounded while the other is not. First, we define what we mean by bounded.

DEFINITION 11.15. *A subset A of the digital plane is **bounded** if there exists $M \in \mathbb{Z}_+$ such that $|m| \leq M$ and $|n| \leq M$ for all $(m, n) \in A$. Otherwise, A is said to be **unbounded**.*

In other words, a subset A of the digital plane is bounded if A is contained in a square centered at the origin in the digital plane.

THEOREM 11.16. *If A is a digital simple closed curve in the digital plane, then one of the components of its complement is bounded and the other is unbounded.*

> *Proof.* Let A be a digital simple closed curve in the digital plane. Since A consists of only finitely many points, A is bounded. Therefore, at least one of the components of its complement must be unbounded. Let $M \in \mathbb{Z}_+$ be such that $|m| \leq M$ and $|n| \leq M$ for all $(m, n) \in A$, and let B and B^* be the sets defined by
>
> $$B = \{(m, n) \in \mathbb{Z}^2 \mid |m| \leq M \text{ and } |n| \leq M\}, \quad B^* = \mathbb{Z}^2 - B.$$

It is straightforward to show that B^* is connected; furthermore, $A \cap B^* = \varnothing$. Thus, B^* is contained in one of the components of the complement of A. It follows that B contains the other component of the complement of A, and therefore that component is bounded.   ∎

If A is a digital simple closed curve in the digital plane, then we call the bounded component of the complement of A the **inside of** A and denote it by $\mathrm{Ins}(A)$, and we call the unbounded component of the complement of A the **outside of** A and denote it by $\mathrm{Out}(A)$. The Digital Jordan Curve Theorem indicates that $A = \partial(\mathrm{Ins}(A))$ and $A = \partial(\mathrm{Out}(A))$ if and only if A is a closed subset of the digital plane.

Now, suppose that we have a digital image—an image displayed on an array of pixels. Associated to the image, we define a partition $\mathcal{P}$ of the visible screen $\mathbb{V}$ in the digital plane. Each set in the partition corresponds to an image region in which every point has the same color. As mentioned previously, we would like to know if it is possible to store the image as a collection of surrounding sets that determine the regions. Thus, we would like to know if there is a collection of digital simple closed curves that determines the sets in the partition $\mathcal{P}$.

The topology on $\mathbb{V}$ is the discrete topology, and therefore $\mathbb{V}$ is totally disconnected. However, in order to formulate our desired theorem, we need to introduce an alternative notion of connectivity for the visible screen.

DEFINITION 11.17.

(i) *Let* $p = (m, n)$ *be a point in the visible screen. The points* $(m - 2, n)$, $(m + 2, n)$, $(m, n - 2)$, *and* $(m, n + 2)$ *are said to be **4-adjacent** to* p.

(ii) *A set* C *in the visible screen is said to be **4-connected** if for every pair of points* $p, q \in C$ *there exists a sequence of points in* C, $p = p_1, p_2, \dots, p_n = q$, *such that* p_{i+1} *is 4-adjacent to* p_i *for all* $i = 1, \dots, n - 1$.

Given a point in the visible screen, the points that are 4-adjacent to it are the four points in the visible screen that are directly next to it, either horizontally or vertically. In Figure 11.19 we show two sets, A and B, in the visible screen. The set A is 4-connected, while the set B is not.

Although we do not need it here, we can similarly define 8-adjacent and 8-connected. In 8-adjacency, all eight open points surrounding a particular open point are considered to be 8-adjacent to it. The notions of 4- and 8-connectivity are important in the formulation of visible-screen versions of topological properties. These connectivity properties played a major role in the development of the field of digital topology by Azriel Rosenfeld. (See [Ros], for example.)

The next definition describes the process by which an image in the visible screen is converted to a collection of surrounding sets in the digital plane.
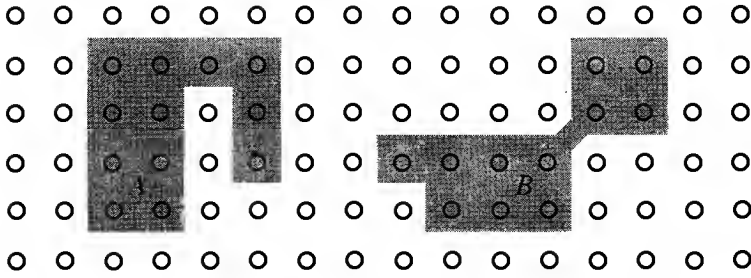
FIGURE 11.19: The set A is 4-connected, but the set B is not.

DEFINITION 11.18. *Let $\mathcal{P}$ be a partition of the visible screen into 4-connected subsets, only one of which is unbounded. The subset of the digital plane defined by*

$$S_{\mathcal{P}} = \bigcup_{D \in \mathcal{P}} \partial(Cl(D))$$

*is called the **cartoon determined by** $\mathcal{P}$.*

Given $\mathcal{P}$ as in the definition, the cartoon determined by $\mathcal{P}$ is a collection, $S_{\mathcal{P}}$, of sets that are obtained by taking (in the digital plane) the boundary of the closure of each set in the partition. In [Kha3] it is shown that $S_{\mathcal{P}}$ is a union of digital simple closed curves in $\mathbb{Z}^2 - \mathbb{V}$, the complement of the visible screen in the digital plane. Thus, we can regard the cartoon determined by $\mathcal{P}$ as a collection of surrounding sets. The following theorem indicates, as desired, that the partition $\mathcal{P}$ can be recovered from the cartoon it determines:

THEOREM 11.19. *Let $\mathcal{P}$ be a partition of the visible screen into 4-connected subsets, at most one of which is unbounded. Let $S_{\mathcal{P}}$ be the cartoon determined by $\mathcal{P}$. Then, the collection $\mathcal{P}^*$, of subsets of $\mathbb{V}$, given by*

$$\mathcal{P}^* = \{C \cap \mathbb{V} \mid C \text{ is a component of } \mathbb{Z}^2 - S_{\mathcal{P}}\},$$

is a partition of the visible screen, and $\mathcal{P}^* = \mathcal{P}$.

We do not provide a proof of Theorem 11.19; it follows from results presented in [Kha3].

In Theorem 11.19 it is necessary to assume that the sets in the partition $\mathcal{P}$ are 4-connected in order to have the partitions $\mathcal{P}^*$ and $\mathcal{P}$ coincide. For example, if the set B in Figure 11.19 is one of the sets in the partition $\mathcal{P}$, then the two 4-connected "components" of B would be separate sets in $\mathcal{P}^*$.

Summarizing, we have the following process for storing and recovering a digital image:

    (i) Given a digital image, define a partition of the visible screen by having each set in the partition correspond to a region in the image with a fixed color.

(ii) In order to store the image, construct the cartoon determined by the partition. (We can also store the color information by indicating which side of each digital simple closed curve in the cartoon has which color.)

(iii) Recover the partition by taking the intersection of the visible screen with each component of the complement of the cartoon in the digital plane.

The challenge of digitally representing the real world, which we usually regard as connected and continuous, requires having a means to "discretize" images into digital versions, while simultaneously having a connectivity structure to maintain real-world topological relationships. The digital plane provides an effective model for this process; it is composed of the totally disconnected, open, and dense visible screen that is used for the representation of images, along with the closed points and mixed points that comprise the invisible structure that provides connectivity.

Euclidean space has been a primary venue for the discipline of topology, but as we continue to rely heavily on digital processes for communicating, presenting, and interpreting information, it will be important to include digital spaces among those that we investigate.

Exercises for Section 11.3

11.17. Given two points in the digital plane, show that there is a continuous function $f : [0, 1] \to \mathbb{Z}^2$ that traces a route traveling vertically, horizontally, or both, from one point to the other.

11.18. Determine the boundary of each component of the complement of the digital simple closed curve A in Example 11.2.

11.19. (a) Consider the four six-point subsets of the digital plane shown in Figure 11.20. Illustrate a basis for the subspace topology on each, and prove that none of the sets is a digital simple closed curve.
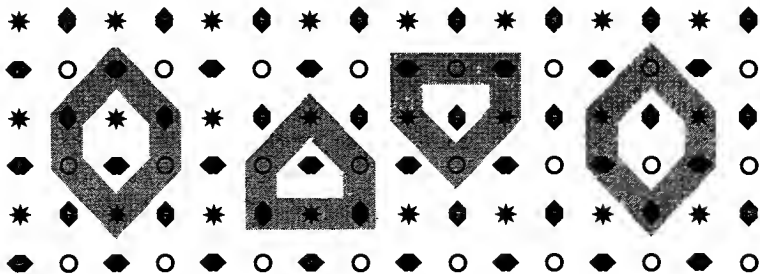
FIGURE 11.20: None of these six-point sets is a digital simple closed curve.

(b) Prove that there are no six-point digital simple closed curves in the digital plane.

(c) Prove that for every even integer $n \geq 8$ there is an n-point digital simple closed curve in the digital plane.

(d) Show that there is a six-point digital simple closed curve in digital 3-space—the space $\mathbb{Z} \times \mathbb{Z} \times \mathbb{Z}$ obtained by taking the product of three digital lines.

11.20. (a) Show that the eight-point set surrounding an open point in the digital plane is a digital simple closed curve.

(b) Show that the eight-point set surrounding a closed point in the digital plane is a digital simple closed curve.

(c) Show that the eight-point set surrounding a mixed point in the digital plane is not a digital simple closed curve. Does such a set separate the digital plane?

11.21. (a) Prove that if S is a simple closed curve in the plane, $\mathbb{R}^2$, then S is a closed subset of the plane.

(b) Why does your proof in (a) fail to carry over to show that if A is a digital simple closed curve in the digital plane, then A is a closed subset of the digital plane?

11.22. Prove Theorem 11.14: Let A be a digital simple closed curve in the digital plane. Then A is a closed subset of the digital plane if and only if A does not contain any open points.

11.23. Let A be an $n \times n$ array of open points in the digital plane. Prove that $\partial(\mathrm{Cl}(A))$ is a digital simple closed curve consisting of $8n$ points.

Chapter 12

Knots

In this chapter, we present the theory of knots. Recall that a knot is an embedding $f : S^1 \to \mathbb{R}^3$ of a circle into $\mathbb{R}^3$. We consider two knots to be equivalent if one can be deformed into the other without passing through itself, as in Figure 12.1. Two knots that are equivalent are said to be of the same knot type. We give a rigorous.definition of this equivalence shortly.
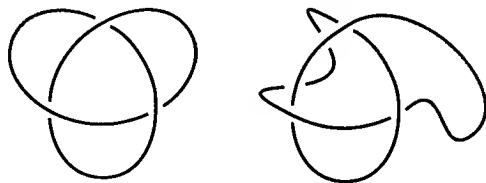
FIGURE 12.1: These two knots are equivalent.

Knot theory was born in the mid-nineteenth century. Carl Friedrich Gauss (1777–1855) and his student Johann Listing (1808–1882) both contributed to the early study of knot theory. Listing's work, "Vorstudien zur Topologie" (where the term "topology" was first used), included a discussion about the problem of classifying knots.

Throughout most of the nineteenth century scientists believed that there was an invisible substance called ether that pervaded all of space. The Scottish physicist William Thomson (1824–1907), also known as Lord Kelvin, posited that atoms were simply knotted vortices in the ether (like the smoke rings one can blow into the air). Each distinct knot was thought to correspond to a distinct element.

Motivated by Thompson's conjecture, James Clerk Maxwell (1831–1879) and Peter Guthrie Tait (1831–1901) studied properties of knots, with Tait working on the development of a catalog of different knots. Ultimately, the list of knots did not correlate well with the list of elements. And, with the Michelson–Morley experiment (1887), it became apparent that there was no ether, and thus a death blow was dealt to this model of the atom. But by then, there was sufficient interest in cataloging knots, solely from a mathematical perspective, that work continued in this direction.

Following the development of the foundations of topology in the early twentieth century, mathematicians were able to formalize the theory of knots. Since then, many mathematical tools have been developed and applied to knot theory, one of the most active areas of topology.

The most fundamental question in knot theory is how to tell whether or not two knots are equivalent. Given pictures of two knots, we wish to determine whether or not it is possible, when the knots are tied in a string, to rearrange one of the knots to look like the other.

One way to prove that two knots are not equivalent is to use quantities, called invariants, that we associate to knots. Invariants depend only on the type of the knot, and not on a particular picture of it. When we have two knots that have different values of an invariant, we know that the knots are distinct.

In Section 12.1, we introduce isotopy and use it to give a rigorous definition of knot equivalence. In Section 12.2, we introduce the Reidemeister moves, which are moves we perform on a knot projection (a picture of a knot) that preserve the type of the knot. Reidemeister moves provide us with a straightforward means for proving that the invariants we define actually are independent of the type of the knot. In this section, we also introduce linking number, an invariant for distinguishing so-called links, which are configurations made up of a collection of knots. In Section 12.3, we introduce polynomial invariants for knots, an innovation dating from the 1920s and 1930s, but which expanded dramatically in the 1980s. In the last section, we introduce some recent applications of knots to the study of DNA and to synthetic chemistry.

12.1   Isotopy and Knots

In Section 9.1, we introduced the concept of a homotopy $F : X \times I \to Y$ between two continuous functions $f : X \to Y$ and $g : X \to Y$ to be a continuous function $F(x, t)$ such that $F(x, 0) = f(x)$ and $F(x, 1) = g(x)$. As the coordinate t varies from 0 to 1, the function F transforms f into g. In knot theory, we are interested in a particular type of homotopy.

DEFINITION 12.1. *A homotopy* $F : X \times I \to Y$ *is called an* **isotopy** *if* $F|_{X \times \{t\}}$ *is a homeomorphism for all t in I.*

In an isotopy $F : X \times I \to Y$, as t varies, we think of the isotopy as a one-parameter family of homeomorphisms $F|_{X \times \{t\}}$ defined on X.

EXAMPLE 12.1. Define $F : \mathbb{R}^2 \times I \to \mathbb{R}^2$ by $F(x, t) = (t + 1)x$. When $t = 0$, F is just the identity map on $\mathbb{R}^2$. But as t grows from 0 to 1, each vector in $\mathbb{R}^2$ beginning at the origin is stretched in length until the end of the isotopy, when they are all twice as long as they originally were. (See Figure 12.2.)
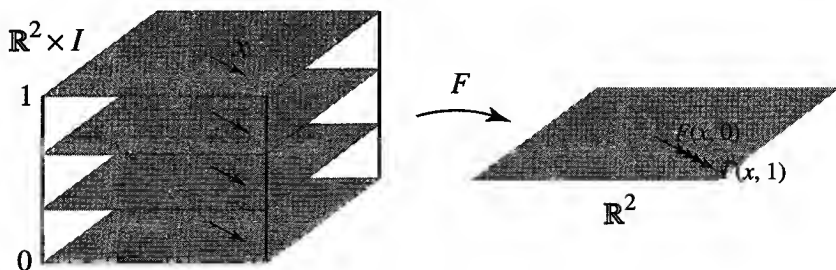
FIGURE 12.2: The isotopy F stretches each vector to twice its length.

We can think of an isotopy as a deformation of a space X over time that does not change the topology of X. We are particularly interested in whether or not two embeddings of a space Y in another space X can be deformed to each other by an isotopy of the space X containing them. In this way, we define what it means for two embeddings to be equivalent.

DEFINITION 12.2. *If $f : Y \to X$ and $g : Y \to X$ are embeddings of Y into X, then we say that f and g are **ambient isotopic** if there is an isotopy $F : X \times I \to X$ such that $F(x, 0) = x$ for all $x \in X$ and $F(f(y), 1) = g(y)$ for all $y \in Y$. The space X is called the **ambient space** and the function F is called an **ambient isotopy**.*

If $F : X \times I \to X$ is an ambient isotopy, then when restricted to $t = 0$, F is just the identity function on X, which fixes all points in X. But when restricted to $t = 1$, F sends the image of a point under f to exactly the same place as the image of the point under g. The ambient isotopy deforms the entire space X in a continious manner, so that by the end the function f has been deformed to the function g.

Notice also that although the conditions $F(x, 0) = x$ and $F(f(y), 1) = g(y)$ appear quite different, the first condition does imply that $F(f(y), 0) = f(y)$. So we can think of these conditions as implying that at time $t = 0$, F sends $f(y)$ to $f(y)$ and at time $t = 1$, F sends $f(y)$ to $g(y)$.

Although complicated to state, the definition of an ambient isotopy captures the true spirit of rubber-sheet geometry. The image of the first embedding is slowly transformed to the image of the second as the ambient space is deformed. Notice that the image of the first embedding is, by definition, homeomorphic to the second. At each stage, a homeomorphic copy of Y appears in the ambient space, and we watch a movie of it as it is transformed from the initial version to the final version. (See Figure 12.3.) Ambient isotopy defines an equivalence relation between embeddings of a space Y into a space X. (See Exercise 12.3.)
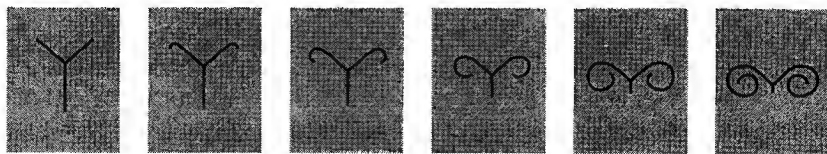
FIGURE 12.3: An isotopy resembles a movie of the deformation.

EXAMPLE 12.2. Define $f : S^1 \to \mathbb{R}^2$ by $f(\theta) = (\cos(\theta), \sin(\theta))$. This embeds the circle onto a copy of itself in $\mathbb{R}^2$. Define $g(\theta) = (2\cos(\theta), 2\sin(\theta))$. This embeds the circle as the circle of radius 2 in $\mathbb{R}^2$, centered at the origin. It seems clear that we can deform the first embedding to the second embedding. (See Figure 12.4.) But what is an explicit ambient isotopy? We can use the isotopy from the previous example. That is, take $F : \mathbb{R}^2 \times I \to \mathbb{R}^2$ given by $F(x, t) = (t + 1)x$. This is an isotopy since for a fixed value t in I, $F|_{\mathbb{R}^2 \times \{t\}}$ is a homeomorphism of the plane to itself.

Moreover, when $t = 0$, $F(x, 0) = x$, so it is the identity map. And finally, when we restrict to $t = 1$, we have that

$$F(f(\theta), 1) = (1 + 1)f(\theta) = 2f(\theta) = g(\theta).$$

Therefore the two embeddings of the circle are, in fact, ambient isotopic.
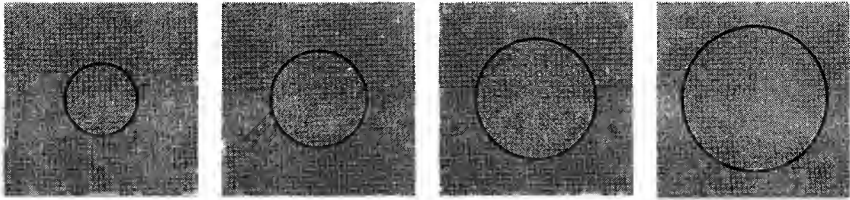
FIGURE 12.4: An isotopy deforming the circle of radius 1 to the circle of radius 2.

EXAMPLE 12.3. There is an ambient isotopy $F : \mathbb{R}^3 \times I \to \mathbb{R}^3$ that takes the sphere S^2 and deforms it to the embedded sphere S depicted in Figure 12.5.
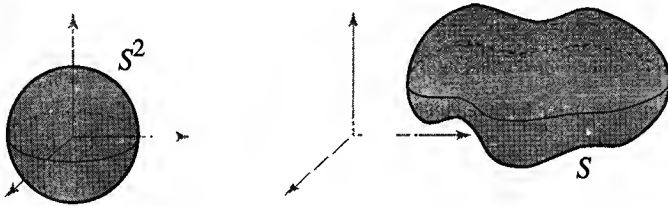
FIGURE 12.5: The sphere S^2 can be deformed to the embedded sphere S by an ambient isotopy.

The function $F|_{\mathbb{R}^3 \times \{0\}}$ is the identity map, which is a homeomorphism. The function $F|_{\mathbb{R}^3 \times \{1\}}$ sends the sphere to the embedded sphere S. We do not give an explicit formula for F. Often it is difficult or impossible to determine a formula for a specific ambient isotopy, although it is apparent from pictures that one exists.

We now wish to apply the concept of ambient isotopy to knots. In order to avoid certain pathological situations, we assume that the knot can be realized by a finite number of sticks glued end-to-end, as in Figure 12.6. We call this a **polygonal knot**. We call the sticks **edges** of the polygonal knot, any adjacent pair of which meet at a **vertex**.

Every smooth knot can be approximated arbitrarily closely by a polygonal knot by using many short sticks, so there is no loss in considering polygonal knots rather than smooth knots.
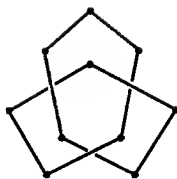
FIGURE 12.6: A knot made from sticks.

We use finitely many sticks so that we can avoid so-called wildly embedded knots, such as the one appearing in Figure 12.7. These knots have strange behavior that belies the intuition we are attempting to capture with knots.
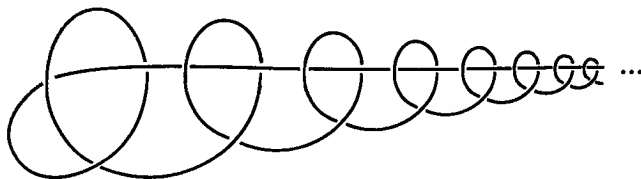
FIGURE 12.7: A wildly embedded knot, which we exclude from consideration.

Again, the mathematical field of knot theory investigates the various embeddings of the circle S^1 into $\mathbb{R}^3$. We would like two distinct knots to be equivalent if we can deform one to the other without passing it through itself. How can we make that relation rigorous?

DEFINITION 12.3. *Two knots* $f, g : S^1 \to \mathbb{R}^3$ *are **equivalent** if they are ambient isotopic. A collection of equivalent knots is called a **knot type**.*

Let $H : \mathbb{R}^3 \times I \to \mathbb{R}^3$ be the ambient isotopy between knots given by the embeddings $f, g : S^1 \to \mathbb{R}^3$. The condition that $H(x, 0) = x$ ensures that H starts as the identity map. The condition that $H(f(x), 1) = g(x)$ ensures that by the time it is done, H has deformed the knot given by f to the knot given by g. The fact that the isotopy must be a homeomorphism for any fixed value of the second coordinate t prevents the knot from passing through itself during the deformation.

EXAMPLE 12.4. All of the knots illustrated in Figure 12.8 are of the same knot type, since any one of them can be deformed to each of the others by ambient isotopies of $\mathbb{R}^3$.
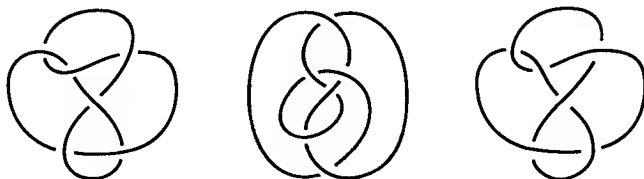
FIGURE 12.8: These knots are of the same knot type.

The reason we use an ambient isotopy of the surrounding space $\mathbb{R}^3$ is to avoid deformations that collapse a part of a knot away, as occurs in Figure 12.9. In that deformation, a part of $\mathbb{R}^3$ is collapsed down to a point at the last stage, and that cannot occur in an ambient isotopy of $\mathbb{R}^3$.
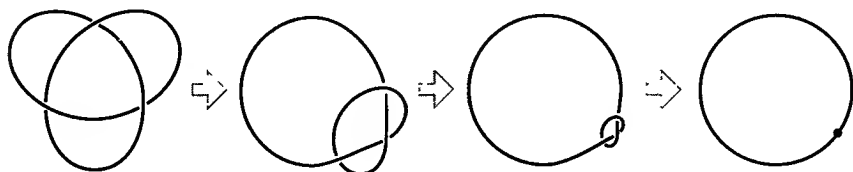
FIGURE 12.9: We do not allow collapsing.

Typically, we picture an ambient isotopy by visualizing how a knot is being deformed through space, keeping in mind that throughout the deformation, the whole ambient space is being deformed along with the knot, and therefore the knot cannot pass through itself. (See Figure 12.10.)
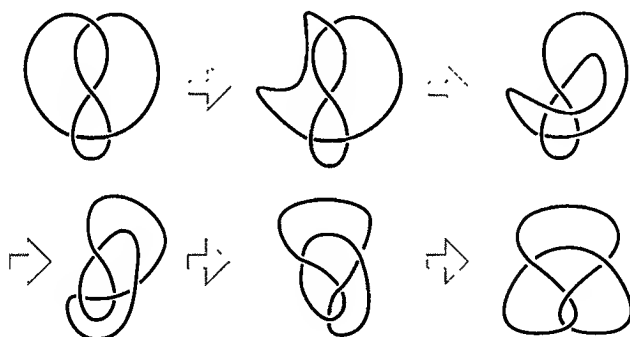
FIGURE 12.10: A sequence of deformations that shows that the initial and final knots are equivalent.

We also assume that the isotopy is what is called **piecewise-linear**. This is to avoid the possibility of passing through a wild knot in the process of deforming one knot to another. A piecewise-linear ambient isotopy from one polygonal knot to another can be realized by a sequence of the following two operations on the knot (see Figure 12.11):

(i) Replace an edge of the knot with two other edges, such that the three edges bound a triangle that intersects the knot along just the first edge.

(ii) Replace two adjacent edges of the knot by one edge, such that the three edges bound a triangle that intersects the knot in exactly the first two edges.

FIGURE 12.11: The operations on a knot in a piecewise-linear ambient isotopy.

We refer to these operations as **triangle moves** on the knot. Each of these moves clearly preserves the knot type, since it can be realized as an ambient isotopy deforming the knot continuously across the triangle to the other edge or edges, as illustrated in Figure 12.12.
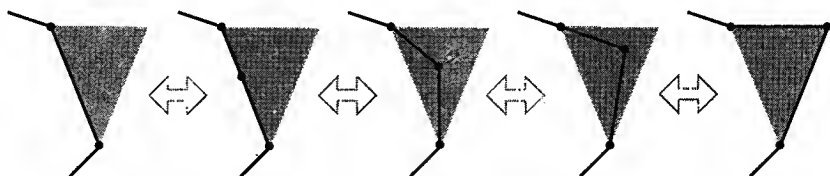
FIGURE 12.12: The triangle moves result from an ambient isotopy.

Although it is worth keeping in mind that all knots should be constructed from finitely many line segments, we do not dwell on this, and we leave it to you to convince yourself that the smooth knots pictured and the ambient isotopies that we describe can be realized by piecewise-linear versions.

It is convenient to use the term knot when referring to the entire equivalence class of a knot. For instance, we say that two knots are distinct when they are not in the same equivalence class. We say that a knot is **trivial** when it is in the same equivalence class as the knot given by the embedding that sends S^1 onto the circle of radius 1, centered at the origin, in the xy-plane in $\mathbb{R}^3$. Only when confusion could arise do we revert to referring to a knot type.

To keep track of knots and the operations we perform on them, we use pictures of knots called knot projections. Given a polygonal knot in 3-space, we can project it to a closed polygonal curve in a plane. The resulting planar curve is called a **projection** of the knot. A projection is called a **regular projection** if

    (i) No point in the projection corresponds to more than two points on the knot;

    (ii) There are only finitely many points in the projection that correspond to two points on the knot. These are called **double points** of the projection;

    (iii) No double point corresponds to a vertex of the knot.

In Figure 12.13, we illustrate several situations causing a projection to fail to be a regular projection.

FIGURE 12.13: Projections that do not result in a regular projection.

If a projection of a knot is not regular, we can move the knot just slightly before projecting to make the projection regular. We call this putting the knot in general position relative to the projection, and although we do not go into the technical details here, it is not hard to convince yourself that we can always do this. For instance, if there is a vertex projecting to a double point, turning the knot in space just slightly will move the vertex off the double point. This slight movement of the knot does not change the knot type. Hence, every knot has a regular projection.

A regular projection of a knot appears as a topological graph in the plane. (See Figure 12.14.) The projected vertices of the knot, together with the double points of the projection, make up the vertices of the topological graph. If the projection of an edge of the knot has no double points in it, then it is an edge of this topological graph. Otherwise, the projected edge of the knot is subdivided into new edges by the vertices corresponding to the double points.
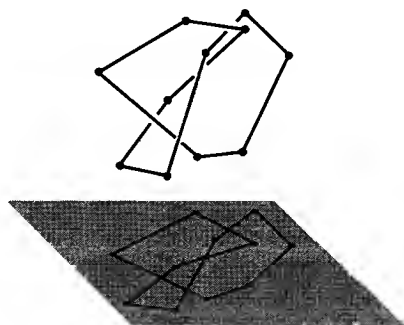
FIGURE 12.14: A regular projection appears as a topological graph in the plane.

A projection of a knot does not carry as much information as does the original knot. At a double point, we do not know which of the two corresponding strands in the knot is closer to the projection plane. However, a **knot projection** is a regular projection that, at each double point, includes an indication of which strand is crossing over the other strand, relative to the projection plane, as in Figure 12.15. Knot projections are also referred to as **diagrams**. Given a knot projection, we can construct a knot in 3-space that is represented by that knot projection. Furthermore, any two such knots that we construct from a given knot projection are equivalent.
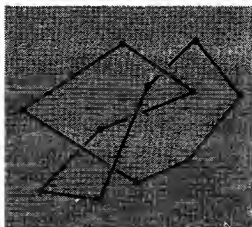
FIGURE 12.15: A knot projection.

We call the double points in a knot projection the **crossings** of the projection.

A **planar isotopy** is a piecewise-linear isotopy of the plane. It deforms a polygonal knot projection to another polygonal knot projection without changing the structure of the associated topological graph in the plane. Although a planar isotopy can stretch and shrink the distances between crossings, it cannot change the number of crossings and which crossings are connected by which edges of the projection.

If we have a planar isotopy from one knot projection to another, then we can construct an appropriate ambient isotopy of $\mathbb{R}^3$ between the corresponding knots, and therefore the knots are equivalent. On the other hand, an ambient isotopy between two knots does not necessarily project to a planar isotopy. For example, in a projection of an ambient isotopy we could see a deformation as illustrated in Figure 12.16. At the third stage we do not have a knot projection.
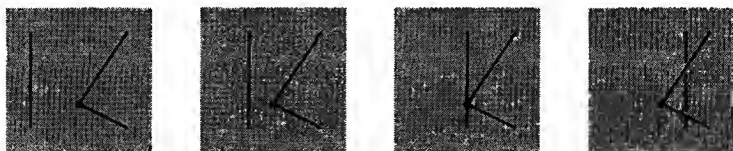
FIGURE 12.16: A projected ambient isotopy that fails to be a planar isotopy.

This deformation plays a fundamental role in our use of knot projections to study knots. It is what is called a Reidemeister move. We investigate Reidemeister moves further in the next section.

The primary goal of knot theory is to find effective means to distinguish between different knot types. In Figure 12.17, we show the collection of knots with up to six crossings. Every knot that has a regular projection with six or fewer crossings is equivalent to one of these knots or to the mirror image of one of these knots. (Note that a knot and its mirror image are not necessarily equivalent. We discuss this further in Section 12.4.)

In the top row in Figure 12.17 the first knot is the trivial knot, the second knot is known as the **trefoil knot**, and the third knot is known as the **figure-eight knot**.

Each of the knots in Figure 12.17 is known to be distinct. If we made one of them out of string, we could never rearrange it to look like any of the others without cutting it open and retying it. But how do we rigorously show these
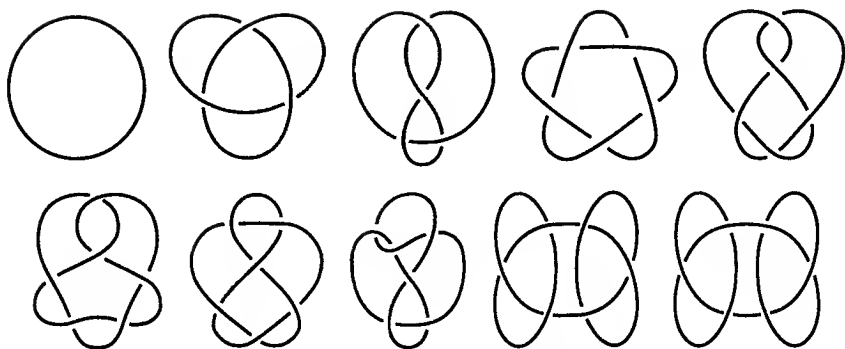
FIGURE 12.17: Distinct knots with six or fewer crossings.

are distinct? We cannot just make one out of string and then try rearranging it to look like another. We might fail for six years, but that would not suffice to conclude that the knots are distinct. How do we know that another five minutes of work would not do the trick?

Instead, we look for so-called invariants that we associate to a given knot and that do not depend on the knot type. There are a variety of different invariants that are used in studying knots, some of them numerical values (see Section 12.2), some of them polynomials (see Section 12.3), and some of them algebraic groups. The important point is that in all these cases, if two knots have different values of a particular invariant, we immediately know they are distinct knots.

But how does one tell if a given candidate invariant actually is an invariant? It is usually too unwieldy to show that ambient isotopy between knots does not change the value of the invariant. There are just too many possible ambient isotopies, and even in the simplest examples, the equations become intractable.

Luckily, we can simplify the task dramatically by using knot projections and applying Reidemeister moves to them. We discuss these ideas in the next section.

Exercises for Section 12.1

12.1. In each case, find an ambient isotopy between the two given embeddings:
    (a) The functions $f, g : I \to \mathbb{R}^2$ given by $f(x) = (x, x)$ and $g(x) = (x^2, x)$
    (b) The functions $f, h : I \to \mathbb{R}^2$ given by $f(x) = (x, x)$ and $h(x) = (0, x)$

12.2. Show that if F is an ambient isotopy between embeddings $f, g : Y \to X$, then f is homotopic to g as functions that map Y to X.

12.3. Show that ambient isotopy defines an equivalence relation on the set of embeddings of a space Y into a space X.

12.4. Sketch a series of pictures showing that the first knot in Figure 12.8 can be deformed to the second, and that the second knot can be deformed to the third.

12.5. Determine which of the knots in Figure 12.18 are equivalent, and conjecture as to which knots are not equivalent. (We can prove that two knots are equivalent by finding a series of pictures that depict the ambient isotopy from one to the other. However, as of yet in this text, we do not have a means of proving that two knots are not equivalent.)
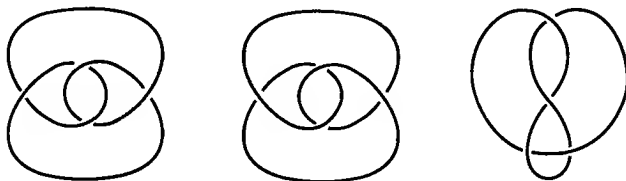
FIGURE 12.18: Which knots are equivalent?

12.6. Most often, knot theorists think of a knot as embedded in S^3 rather than $\mathbb{R}^3$. As we have seen, we can think of S^3 as the one-point compactification $\mathbb{R}^3 \cup \{\infty\}$, so there is just a single-point difference between the two spaces. The advantage to adding that single point to obtain S^3 is that the resulting ambient space is compact, which turns out to be useful. Show that if two knots are equivalent in $\mathbb{R}^3$, they are equivalent in S^3 as well.

12.2 *Reidemeister Moves and Linking Number*

Suppose that two knot projections represent equivalent knots. Then there must be an ambient isotopy from the knot represented by the first projection to the knot represented by the second projection. However, determining the existence or nonexistence of such an isotopy is difficult, since isotopies give us so much leeway. In 1927, Kurt Reidemeister (1893–1971) simplified the task by showing that the existence of an ambient isotopy between knots represented by two different knot projections is equivalent to the existence of a sequence of moves, called Reidemeister moves, that take us from one projection to the other. The Reidemeister moves fall into three types.

The Type I Reidemeister moves allow us to put in or take out a kink in a projection, as in Figure 12.19. Notice that this certainly does not change the knot type.
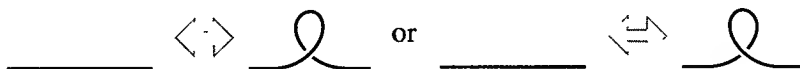
FIGURE 12.19: The Type I Reidemeister moves.

For the Type II Reidemeister moves, we slide one of two adjacent strands under the other, as in Figure 12.20, or, taking the reverse perspective, we slide one strand out from under another, resulting in two adjacent strands. The two moves illustrated in Figure 12.20 are really the same. If we rotate the illustration of the right-hand move by 180 degrees we obtain an illustration that essentially shows the left-hand move. It is clear that the knot type is preserved by the Type II Reidemeister moves.
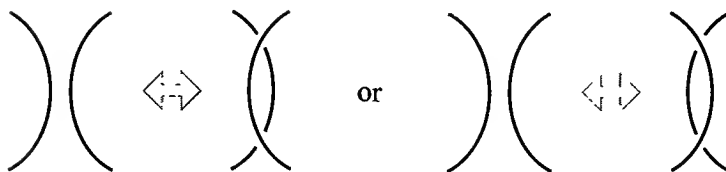
FIGURE 12.20: The Type II Reidemeister moves.

For the Type III Reidemeister moves, a strand is slid past a crossing, over the two strands that make up the crossing, as illustrated in Figure 12.21.

If we examine the moves in Figure 12.21 from the perspective of the bottom strand, then in the illustrated moves the bottom strand moves under a crossing formed by the other two strands. Thus the Type III moves, as shown, include the possibility of moving a strand under a crossing. Furthermore, the illustrated Type III moves also include the possibility of moving a strand past a crossing, between the two strands that make up the crossing.
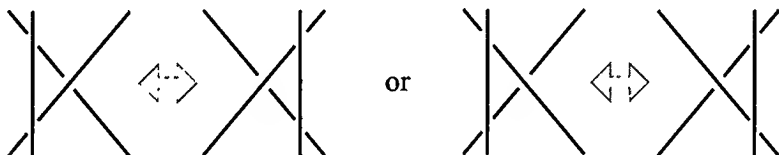
FIGURE 12.21: The Type III Reidemeister moves.

Next, we introduce Reidemeister's Theorem, a significant result, which, as we will subsequently see, enables us to verify that the knot invariants we define are indeed invariant of the knot type.

THEOREM 12.4. *Reidemeister's Theorem.*   *Two knots are equivalent if and only if there is a finite sequence of planar isotopies and Reidemeister moves taking a knot projection of one to a knot projection of the other.*

> ***Proof.*** As mentioned previously, planar isotopies and Reidemeister moves do not change knot type, so it is immediate that if two knot projections are related through a sequence of planar isotopies and Reidemeister moves, then they represent equivalent knots.
>
> Suppose now that two knot projections correspond to equivalent knots. We need to show that there is a sequence of planar isotopies and Reidemeister moves from one to the other.
>
> We assume that both of our knots are polygonal, and therefore the topological graphs associated to the corresponding knot projections are made up of line segments attached at vertices. Since the two knots represented are equivalent, there is a piecewise-linear ambient isotopy from $\mathbb{R}^3$ to $\mathbb{R}^3$ that deforms one knot to the other. Because the isotopy is piecewise-linear, by definition we can realize its impact on the knot by a sequence of triangle moves.

As long as the projection of each stage of the ambient isotopy is a knot projection, then we are simply following along a planar isotopy. But what happens when a stage of the ambient isotopy does not yield a regular projection? It can be shown that there are essentially three ways that this can happen.

Consider the triangle move illustrated in Figure 12.22. At stages (c) and (e) the projection is not a regular projection. As the triangle-move deformation progresses between stages (b) and (d) through (c), the result in the projection is a Type III Reidemeister move. And, as the deformation progresses between (d) and (f) through (e), the result is a Type II Reidemeister move.
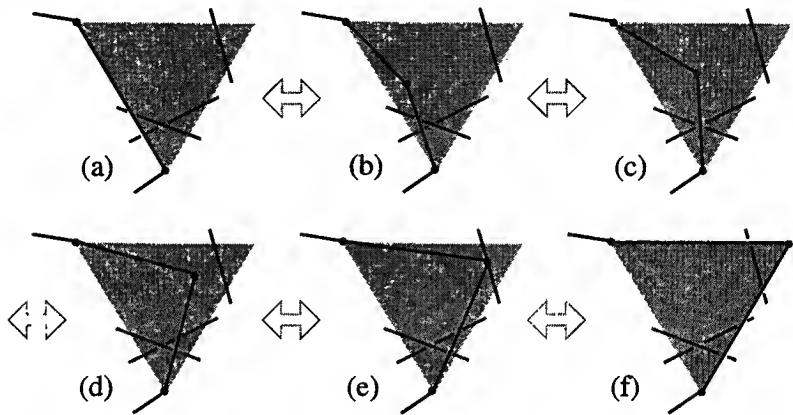
FIGURE 12.22: The Type III and Type II Reidemeister moves occur during a triangle move as illustrated.

In Figure 12.23 we illustrate the third situation where a projection of an ambient isotopy is not a regular projection. With the projection of the edges as shown, in stage (c) we do not have a regular projection. The overall change in the projection, as the deformation goes from stage (a) to stage (e), or vice versa, is a Type I Reidemeister move.
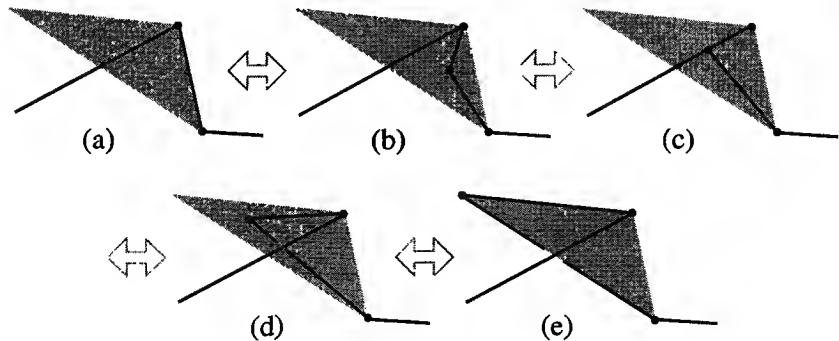
FIGURE 12.23: The Type I moves occur when the edges in the triangle move are arranged as shown.

Thus, when the triangle moves are projected in a diagram they result in a sequence of planar isotopies or Reidemeister moves. Therefore if two knots are equivalent, then we can get from a knot projection of the first to a knot projection of the second by a sequence of planar isotopies and Reidemeister moves.   ∎

In addition to knots, we also consider links.

DEFINITION 12.5. *A **link** is an embedding of a set of circles in $\mathbb{R}^3$. Two links are considered equivalent if one can be deformed to the other via ambient isotopy. Each embedded circle is called a **component** of the link and a link is called an n-**component link** if it has n components.*

Figure 12.24 depicts two of the more well-known links, the Whitehead link and the Borromean rings. The first is a two-component link and the second is a three-component link. The Borromean rings appeared on the family crest of the Borromea family during the Italian Renaissance.
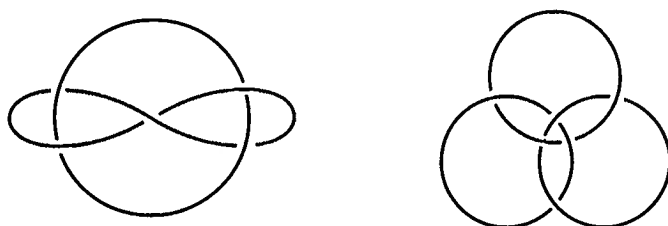
FIGURE 12.24: The Whitehead link and the Borromean rings.

All of the ideas and results presented for knots (regular projections, Reidemeister moves, Reidemeister's Theorem, and so on) carry over here for links. There is an easily computed invariant, called the linking number, that we use to distinguish links from each other. We introduce it next.

Let L be a two-component link. On each component, we pick a direction of travel around the component. We call the direction an **orientation**, and we denote it by placing arrows on the component. In a particular diagram of the link, each crossing appears as one of the two pictures in Figure 12.25.
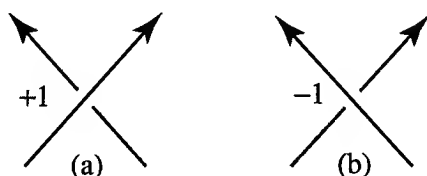
FIGURE 12.25: Possible crossings in a projection of an oriented link.

Label each crossing that occurs between the two different components with a $+1$ or -1, according to whether the crossing matches (a) or (b) in Figure 12.25. We denote the label of a crossing c by $l(c)$.

DEFINITION 12.6. *The **linking number** of a diagram of an oriented two-component link L is given by*

$$lk(L) = \frac{1}{2} \sum_c l(c),$$

where the sum is taken over all of the crossings c involving both of the components in the link.

EXAMPLE 12.5. We determine the linking number of the diagram of the oriented link L that appears in Figure 12.26.
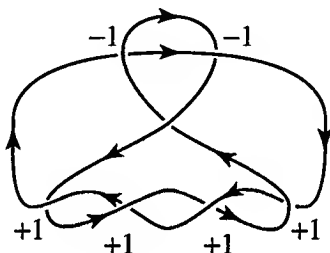
FIGURE 12.26: The oriented link L with its crossings labeled.

There are four crossings with label $+1$ and two crossings with label -1. We do not label the crossings that occur between a component and itself. So, $lk(L) = \frac{1}{2}(+1 + 1 + 1 + 1 - 1 - 1) = 1$.

Linking number is defined for a link projection associated to a given oriented link. The next theorem shows that linking number is actually well defined for the oriented link itself, independent of the choice of a link projection. Given that, the theorem implies that linking number is an invariant for oriented links.

THEOREM 12.7. *If two oriented links are equivalent, then all of their diagrams have the same linking number.*

> **Proof.** We must show that if we have two link projections of equivalent oriented links, then the linking number computed from each one is the same. As with knot projections, if two links are equivalent, then their link projections are related through a sequence of planar isotopies and Reidemeister moves. The planar isotopies do not change the crossings and hence do not affect the linking number. Thus, it is enough to show that the linking number is unaffected by Reidemeister moves.

Suppose we do a Type I Reidemeister move. It either eliminates or creates a crossing between a single component and itself. However, it does not change the crossings between the components. So the linking number is unaffected by a Type I move.

Suppose we slide a strand under an adjacent strand via a Type II Reidemeister move. If the two strands are from the same component, this does not change the linking number. If the two strands are from different components, we create a $+1$ crossing and a -1 crossing. Since in the linking number, the $+1$s and -1s are summed, this leaves the linking number unaffected. Similarly, if we eliminate two crossings between distinct components via a Type II move, we eliminate a $+1$ crossing and a -1 crossing, and therefore the linking number remains unchanged.

Finally, if we do a Type III Reidemeister move, the linking number is unchanged. We simply move the positions of the labeled crossings, so we do not change the overall sum of the labels.

Thus, the Reidemeister moves leave the linking number unchanged, and therefore all diagrams of equivalent oriented links have the same linking number. The linking number is an invariant of oriented links. ∎

EXAMPLE 12.6. Consider the links in Figure 12.27. The link T is called the **trivial link of two components**. It has linking number 0. The link in the middle, M, has linking number $+2$ or -2 depending on the choice of orientations. Thus, link M is not trivial.

The linking number of the Whitehead link, W, is also 0. Hence, linking number does not distinguish W from the trivial link, but it does distinguish W from M.
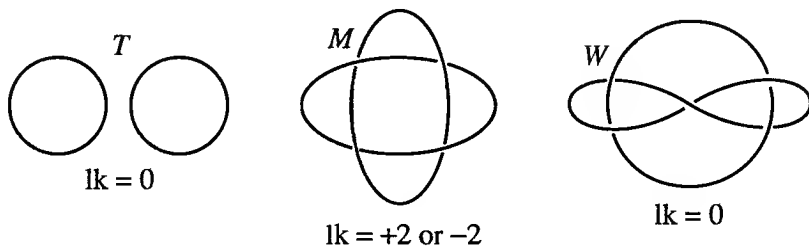
FIGURE 12.27: The linking number distinguishes T and W from M, but not T from W.

Exercises for Section 12.2

12.7. Determine a sequence of Reidemeister moves that would take us from the first projection of the knot depicted in Figure 12.28 to the second projection.
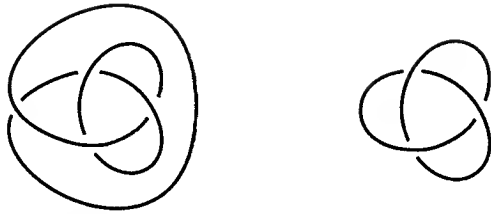
FIGURE 12.28: How does the first projection transform to the second?

12.8. Use linking number to distinguish between the two links shown in Figure 12.29.
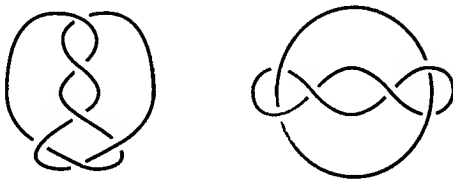
FIGURE 12.29: Show that these links are not equivalent.

12.9. Use linking number to show that the link appearing in Figure 12.30 is not the trivial link of two components.
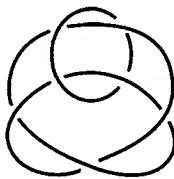
FIGURE 12.30: Show that this link is not trivial.

12.10. Use linking number to prove that the two links shown in Figure 4.20 are not equivalent, and then use that result to argue that there is no ambient isotopy between the two embeddings of the annulus illustrated in Figure 4.19.

12.11. Prove that the two three-component links in Figure 12.31 are not equivalent.
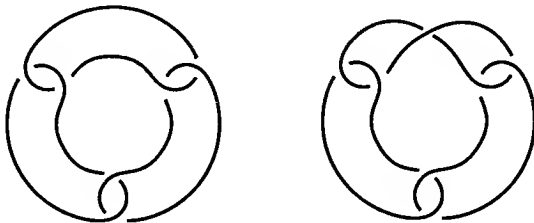
FIGURE 12.31: Show that these two links are not equivalent.

12.3 *Polynomials of Knots*

In his 1928 paper, "Topological invariants of knots and links," J. W. Alexander (1888–1971) introduced a polynomial invariant for knots. This invariant made it possible to distinguish between knots by showing that they have different associated polynomials. However, the Alexander polynomial is limited in the distinctions it allows—there are different knots that have the same Alexander polynomial. In the 1980s, Vaughan Jones discovered a new polynomial invariant, now known as the Jones polynomial. This resulted in a dramatic increase in the number of knots that could be distinguished. Since the introduction of the Jones polynomial, a variety of new polynomial invariants have been discovered. Here, we introduce the Jones polynomial using the approach of Louis Kauffman.

Suppose that we want to create a polynomial invariant for knots and links. We assume that it is a Laurent polynomial, meaning the variable can appear in the polynomial with either positive or negative integer exponents. We would like to be able to compute the polynomial from a given knot projection. However, for it to be an invariant, it cannot depend on the particular projection. We need to know that it is unaffected by changes in the knot projection.

In the previous section, we saw that any two distinct projections of a given knot are related by a sequence of Reidemeister moves. So if we can create a polynomial from a knot projection such that it is unaffected by Reidemeister moves, it will be the same for all knot projections. We will have created a polynomial invariant for knots.

First, let us assume that we are creating a polynomial in three variables A, B, and C that can be calculated from a knot or link projection. Initially, we will not assume that it is the same for all projections of a given knot. Ultimately, we will reduce the polynomial to a polynomial in one variable A as we work toward defining a polynomial invariant from it. We call the resulting polynomial a **bracket polynomial**, and we denote the bracket polynomial of a given knot or link projection P by $<P>$. For simplicity, in the remainder of the section we say "projection" where we mean "knot or link projection."

We assume that three rules are satisfied by the polynomial.

Rule 1: $<\bigcirc> = 1$ where $\bigcirc$ is a trivial-knot projection with no crossings.

This says that a knot projection with no crossings has polynomial 1. This is what is called a normalization condition.

Rule 2: $<P \cup \bigcirc> = C<P>$.

This rule indicates that adding an extra, unlinked trivial component $\bigcirc$ to a projection P multiplies the polynomial of P by C.

Rule 3: $<\asymp> = A<)(> + B<\smile\frown>$,
$<\times> = A<\smile\frown> + B<)(>$.

Rule 3 says that we can obtain the polynomial associated to a projection P in terms of the polynomials associated to two new projections, each with one

fewer crossing than P. This is called a **skein relation**. There are two versions of this relation, but they are essentially the same.

The first version of Rule 3 indicates that given a projection, if we take a crossing in it that is arranged as shown and we cut through the crossing vertically to obtain a new projection and we cut through the crossing horizontally to obtain another new projection, then the bracket polynomials of the three projections are related as in the equation. The second version of Rule 3 can be interpreted similarly, but note that if we rotate each diagram in the second version by 90 degrees, we obtain the first version, and therefore, as already indicated, the two versions of the relation really are the same. How the skein relation works should become clear as we employ it in a number of situations hereafter.

Given these rules, we want to make choices for B and C so that the resultant polynomial is an invariant. In particular, we want it to be unaffected by the Reidemeister moves. Let us start with the Type II Reidemeister move. Applying the bracket polynomial rules, we obtain the following:

$$
\begin{aligned}
<\!\!\big)(\!\!> &= A<\!\!\overset{\frown}{\smile}\!\!> + B<\!\!\big)(\!\!> \\
&= A[A<\!\!\smile\!\!> + B<\!\!\overset{\frown}{\smile}\!\!>] + B[A<\!\!)(\!\!> + B<\!\!\smile\!\!>] \\
&= AA<\!\!\smile\!\!> + ABC<\!\!\smile\!\!> + BA<\!\!)(\!\!> + BB<\!\!\smile\!\!> \\
&= [A^2 + ABC + B^2]<\!\!\smile\!\!> + BA<\!\!)(\!\!>.
\end{aligned}
$$

To have the bracket polynomial unaffected by Type II Reidemeister moves, we want $<\!\!)(\!\!> = <\!\!)(\!\!>$. From the relationship we just derived, it follows that we need to have $BA = 1$ and $A^2 + ABC + B^2 = 0$. Therefore we set $B = A^{-1}$ and $C = -A^2 - A^{-2}$. The original three rules then become

Rule 1: $<\!\!\bigcirc\!\!> = 1.$

Rule 2: $<\!\!P \cup \bigcirc\!\!> = [-A^2 - A^{-2}]<\!\!P\!\!>.$

Rule 3: $<\!\!\times\!\!> = A<\!\!)(\!\!> + A^{-1}<\!\!\smile\!\!>,$
$<\!\!\times\!\!> = A<\!\!\smile\!\!> + A^{-1}<\!\!)(\!\!>.$

With these new rules, we know that the resulting polynomial will be unchanged by Type II Reidemeister moves.

Next, we look at the bracket polynomials of a few specific projections. Following that, we return to examining how the bracket polynomial is affected by Reidemeister moves.

EXAMPLE 12.7. The following bracket polynomials are not difficult to verify using Rules 1–3. We ask you to do so in Exercise 12.12.

$$
<\!\!\bigcirc\bigcirc\!\!> = A^2 - A^{-2}
$$
$$
<\!\!\bigcirc\!\!> = -A^{-3} \qquad <\!\!\bigcirc\!\!> = -A^3
$$
$$
<\!\!\bigcirc\!\!> = -A^{-3} \qquad <\!\!\bigcirc\!\!> = -A^3
$$

EXAMPLE 12.8. Here we compute the bracket polynomial of the projection ⟨◉⟩ of the trefoil knot. Rule 3 is used to obtain the first two equalities shown, and then Rule 2 and a little algebra are used to obtain the third equality:

$$\langle \mathcal{G} \rangle = A\langle \mathcal{U} \rangle + A^{-1}\langle \mathcal{G} \rangle$$
$$= A[A\langle \mathcal{U} \rangle + A^{-1}\langle \mathcal{U} \rangle] + A^{-1}[A\langle \mathcal{G} \rangle + A^{-1}\langle \mathcal{G} \rangle]$$
$$= A^2[-A^2 - A^{-2}]\langle \mathcal{OO} \rangle + 2\langle \mathcal{OO} \rangle + A^{-2}\langle \mathcal{O} \rangle.$$

Now, if we substitute in values from the bracket polynomials in Example 12.7 and then simplify, we obtain

$$\langle \mathcal{G} \rangle = A^2[-A^2 - A^{-2}](-A^3) + 2(-A^3) + (A^{-2})(-A^{-3})$$
$$= A^7 - A^3 - A^{-5}.$$

Therefore the bracket polynomial of the projection ⟨◉⟩ is $A^7 - A^3 - A^{-5}$.

We now return to our investigation of the impact of Reidemeister moves on the bracket polynomial. What happens under Type III Reidemeister moves? For the first form of the Type III move, we obtain the following:

$$\langle \mathcal{K} \rangle = A\langle \mathcal{)(} \rangle + A^{-1}\langle \mathcal{K} \rangle$$
$$= A\langle \mathcal{)(} \rangle + A^{-1}\langle \mathcal{K} \rangle$$
$$= \langle \mathcal{X} \rangle.$$

The first and third equalities hold by the skein relation, and the second equality holds by the fact that the bracket polynomial is unchanged by Type II moves. Conveniently enough, the bracket polynomial is unaffected by a Type III move of the first form, given that it is unaffected by a Type II move with our choice of B and C. It can similarly be shown that the bracket polynomial is unaffected by a Type III move of the second form.

Finally, we must consider the impact of Type I Reidemeister moves. Applying the skein relation and our rules, we obtain the following:

$$\langle \mathcal{S} \rangle = A\langle \mathcal{\Lambda} \rangle + A^{-1}\langle \mathcal{o} \rangle$$
$$= A\langle - \rangle + A^{-1}[-A^2 - A^{-2}]\langle - \rangle$$
$$= -A^{-3}\langle - \rangle.$$

Moves of Type I have an impact on the polynomial, creating a serious problem. We do not obtain the needed equality, $\langle \mathcal{S} \rangle = \langle - \rangle$, for the bracket polynomial to be invariant under the first form of the Type I Reidemeister move. A similar situation occurs with the second form of the Type I Reidemeister move.

To deal with this issue, we need one other concept, called the writhe. Using it, we will be able to define a knot and link invariant from the bracket polynomial. The writhe is defined like linking number, only we consider every crossing in a projection, not just those involving different components of a link.

DEFINITION 12.8. *Given a projection P of an oriented link, the* **writhe** *of P, denoted w(P) is the sum of the labels, +1 or −1, at all of the crossings in P.*

EXAMPLE 12.9. The writhe of the first projection in Figure 12.32 is 0, while the writhe of the second one is 7.
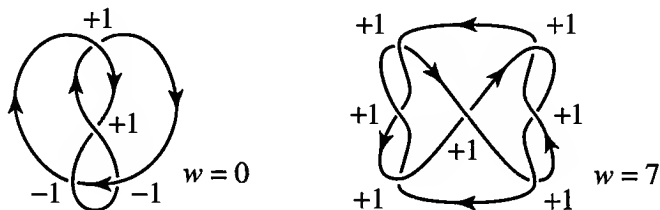
FIGURE 12.32: Computing the writhe of the given projections.

Notice that when we perform a Type I Reidemeister move on a projection, as in Figure 12.33, the writhe either decreases by 1 or increases by 1.
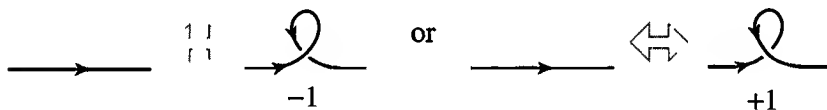
FIGURE 12 33: A Type I move decreases or increases the writhe by 1.

At this point we have defined the bracket polynomial for projections. It is unchanged by Type II and Type III Reidemeister moves, but it changes under Type I moves. Therefore the bracket polynomial is not an invariant for knots and links since it depends on the particular projection. We have also defined the writhe, a property of projections of oriented links. We use the writhe with the bracket polynomial to obtain a new polynomial defined for projections of oriented links. As we show in Theorem 12.10, this polynomial does not depend on the particular projection of a knot or link and therefore is a polynomial invariant.

DEFINITION 12.9. *The* **Kauffman X polynomial** *of a projection P of an oriented link is defined to be the polynomial* $X(P) = (-A^3)^{-w(P)} <P>$.

THEOREM 12.10. *The Kauffman X polynomial is an invariant for knots and for oriented links.*

This theorem indicates that the X polynomial is an invariant for knots, whether they are oriented or not, and for links that are oriented. The reason for this difference is that the writhe of a projection of an oriented knot is not changed if the orientation of the knot is changed, but the writhe of an oriented link can change if the orientation on one of the link components is changed.

Proof. To prove the theorem we need to show that the X polynomial is the same for all projections of a given knot or oriented link. Therefore we need to show that it is unaffected by the Reidemeister moves.

First consider the Type II Reidemeister move. We have already seen that the bracket polynomial is unaffected by Type II moves. The writhe is also unchanged by Type II moves because such moves either add two crossings of opposite sign or eliminate two crossings of opposite sign. Thus the X polynomial is unaffected by Type II Reidemeister moves.

Next, consider the Type III Reidemeister moves. Here too, we have already seen that the bracket polynomial is not affected by such moves. Furthermore, a Type III move just rearranges crossings, without changing the labels on any of them. Therefore the writhe is unchanged by Type III moves. It follows that the X polynomial is unaffected by Type III Reidemeister moves.

Finally, when we perform a Type I Reidemeister move, putting a kink in our projection as on the left in Figure 12.33, we see that the writhe goes down by one. This results in the polynomial being multiplied by $-A^3$. However, the bracket polynomial is multiplied by a factor of $-A^{-3}$ by that same move, as seen in the relationship $<\!\mathcal{S}\!> = -A^{-3}<\!\!-\!\!>$ previously derived. Therefore the X polynomial is unaffected by the first form of a Type I Reidemeister move. We leave it to you to check that the X polynomial is not affected by the other form of a Type I Reidemeister move.

It follows that the Kauffman X polynomial is an invariant for knots and oriented links.   ∎

EXAMPLE 12.10. We calculate the Kauffman X polynomial for the trefoil knot $\mathcal{G}$. From Example 12.8 we have the bracket polynomial

$$<\!\mathcal{G}\!> = A^7 - A^3 - A^{-5}.$$

The writhe of the projection $\mathcal{G}$ is -3. Therefore the Kauffman X polynomial of the trefoil knot is

$$X(\mathcal{G}) = (-A^3)^3(A^7 - A^3 - A^{-5}) = -A^{16} + A^{12} + A^4.$$

As a result of Example 12.10, we can now prove that there are nontrivial knots. In particular, we have the following theorem:

THEOREM 12.11. *The trefoil knot is not equivalent to the trivial knot.*

Proof. The Kauffman X polynomial of the trivial knot is 1 and therefore is not equal to the Kauffman X polynomial of the trefoil knot. Since the Kauffman X polynomial is an invariant for knots, it follows that the trefoil knot is not equivalent to the trivial knot.   ∎

We have shown that there are nontrivial knots. Not only have we done that, but we also have a means for distinguishing between many knots. In fact, with the Kauffman X polynomial, it is possible to distinguish between all of the knots with six or fewer crossings shown in Figure 12.17.

The Kauffman X polynomial is equivalent to the Jones polynomial, which was first discovered in 1984. The Jones polynomial is defined using a slightly different set of rules than the three we used in defining the bracket polynomial; the result is a polynomial with fractional exponents. The Jones polynomial can be obtained from the Kauffman X polynomial, however, by substituting $t^{-1/4}$ for A.

Exercises for Section 12.3

12.12. Compute the bracket polynomials of each of the projections in Example 12.7, verifying the polynomials presented there.

12.13. **(a)** Compute the bracket polynomial of the trivial link of n components shown on the left in Figure 12.34.

      (b) Compute the bracket polynomial of the twisted trivial-knot projections shown on the right in Figure 12.34.

n components

n crossings
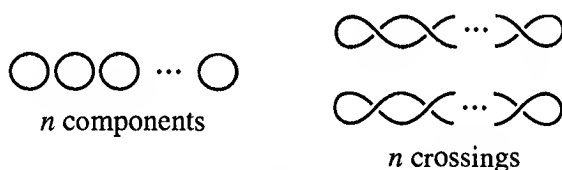
FIGURE 12.34: The trivial link of n components and the twisted trivial-knot projections.

12.14. Compute the writhe of the projections shown in Figure 12.35.
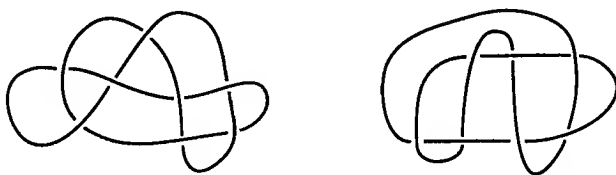
FIGURE 12.35: Compute the writhe of each projection.

12.15. Compute the X polynomial of the mirror image ⊗ of the trefoil knot whose X polynomial we computed in Example 12.10. What does your result imply about the knots ⊗ and ⊗?

12.16. Compute the X polynomial of the figure-eight knot (the third knot in the top row in Figure 12.17).

12.17. Compute the X polynomials of the trivial link of two components and each of the oriented Hopf links shown in Figure 12.36. Conclude that these three oriented links are distinct.
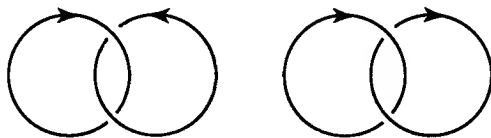
FIGURE 12.36: Oriented Hopf links.

12.18. The original Jones polynomial $V(L)$ of a knot or link L is obtained from the X polynomial by the substitution $A = t^{-1/4}$.
   (a) Find the Jones polynomial of the figure-eight knot.
   (b) Use the skein relation to show that the Jones polynomial satisfies the skein relation $t^{-1}V(L_+) - tV(L_-) + (t^{-1/2} - t^{1/2})V(L_0) = 0$, where L_+, L_-, and L_0 are three projections that are identical except for the portions appearing in Figure 12.37.
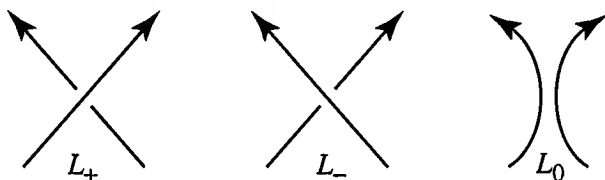
FIGURE 12.37: The projections L_+, L_-, and L_0 differ in only one place, as shown here.

12.4   Applications to Biochemistry and Chemistry

In this section, we consider applications of knot theory to enzyme actions on DNA and to chemical design.

Knots in DNA

We introduced the structure of DNA in Section 5.2, where we indicated that a DNA molecule is made up of two chains of nucleotides that bond together. The two chains are tightly wrapped around one another, not unlike two lengths of twine that have been braided into a thicker rope. Despite this higher-order structure, a DNA molecule—like a rope—can be modeled as a single long, thin strand. In addition, we call a DNA molecule cyclic if one end of the molecule is glued to the other. Cyclic DNA is common, but, much of what we say in this section holds even for noncyclic DNA.

    As we stated in the introduction to the book, the level of DNA compaction in the nucleus of a cell is equivalent to stuffing 200 kilometers of fishing line into a basketball. Problems relating to DNA tangling are therefore quite serious for biological systems. As an example, consider the process of DNA replication. The two nucleotide chains from one DNA molecule separate, and each is used

as a template for construction of a complete, new molecule. Because the templates are heavily intertwined at the start, the two new molecules are born tangled together. For them to be useful, the cell must be able to separate them. How does this occur?

Inside the nucleus of the cell, there are enzymes that allow the DNA to disentangle and tangle. Biochemists would like to understand the action of each of these enzymes. One such enzyme, topoisomerase II, takes two strands of DNA, cuts one open, passes the other through, and then closes up the first. This amounts to a crossing change, as shown in Figure 12.38.
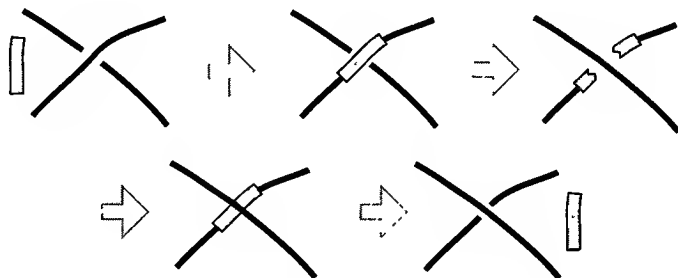
FIGURE 12.38: An enzyme making a crossing change.

We would like to know whether such moves are sufficient to untangle DNA molecules. The following theorem answers that question in the affirmative:

THEOREM 12.12. *Given a knot projection, there is a subset of the crossings that can be switched to obtain a diagram of the trivial knot.*

Proof. Given a knot projection P, we define an algorithm for switching crossings so that the resulting knot projection P' corresponds to the trivial knot.

The Unknotting Algorithm: On P, choose a noncrossing point as a starting point, and choose a direction of travel. (See Figure 12.39.) As we travel around P from the starting point, each time we arrive at a crossing for the first time, change it to an overcrossing relative to our position if it is not already an overcrossing. (When we arrive at a crossing for the second time, it will be an undercrossing relative to our position, and we leave it as is.) Continue this process as we travel all of the way around P, back to the beginning.

The example in Figure 12.39 shows how the Unknotting Algorithm changes crossings (the ones that are shaded) to produce the trivial knot.

Given a knot projection P, let P' be a knot projection obtained by applying the unknotting algorithm to P. We claim that P' is a projection of the trivial knot. To establish this, we construct a knot K in $\mathbb{R}^3$ having knot projection P', then show that K is trivial.

Assume that the knot projection P' lies in the xy-plane. We construct K so that it projects directly down onto P'. Therefore, above each point
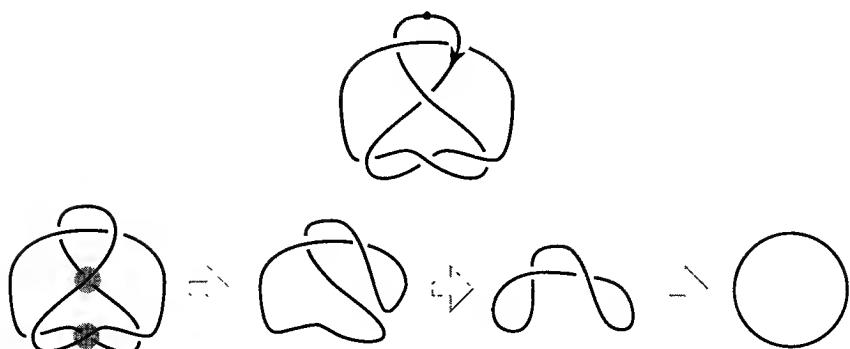
FIGURE 12.39: Changing crossings to unknot a knot.

(x, y) in P', we choose a point (x, y, z) on the knot in 3-space. The points on the knot are chosen so that as we travel around P' in the given direction from the starting point, the z-coordinate decreases continuously until just after our last encounter with a crossing, then increases so that we make our way back up to the first point we chose on the knot. (See Figure 12.40.)
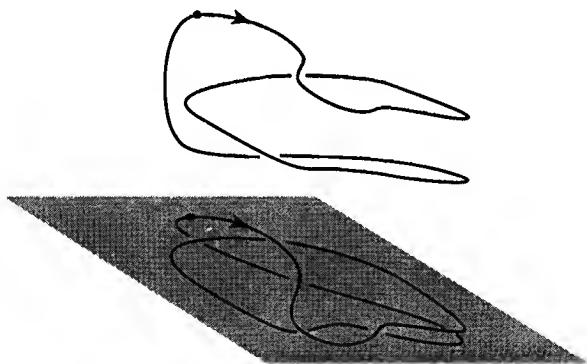
FIGURE 12.40: Building a knot from a knot projection.

Clearly, if we project K directly down to the xy-plane, it projects onto P'. Furthermore, since the z-coordinates are chosen as they are, we have the proper overcrossing/undercrossing relationship at each crossing in the diagram. Finally, with the knot K as constructed, it is not difficult to see that the decreasing-z part of K and the increasing-z part of K can be deformed to separate halves of a circle making up a trivial knot in $\mathbb{R}^3$. (See Figure 12.41.) It follows that K is trivial, and P' is a knot projection of K. ∎
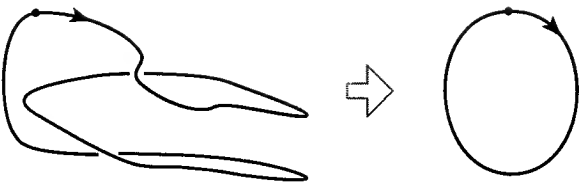
FIGURE 12.41: The knot K deforms to a trivial knot.

As a consequence of Theorem 12.12, every knot has a minimum number of crossing changes needed to unknot it.

DEFINITION 12.13. *The **unknotting number** of a knot K is the minimum number of crossings, over all knot projections of K, that must be changed in order to turn some knot projection of K into one corresponding to the trivial knot.*

EXAMPLE 12.11. The figure-eight knot shown on the left in Figure 12.42 is nontrivial. (See Exercise 12.16.) However, one crossing change turns it into the trivial knot. Therefore its unknotting number is 1.
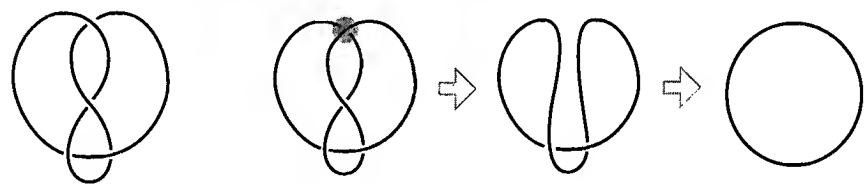
FIGURE 12.42: The figure-eight knot can be made trivial with one crossing change.

EXAMPLE 12.12. The 7_4 knot shown on the left in Figure 12.43 can be unknotted with the two crossing changes shown on the right. However, proving that its unknotting number is 2 requires showing that it does not possess a diagram that, with one crossing change, can be made into a diagram of the trivial knot. Although this can be shown, we do not have the tools necessary to prove it here.
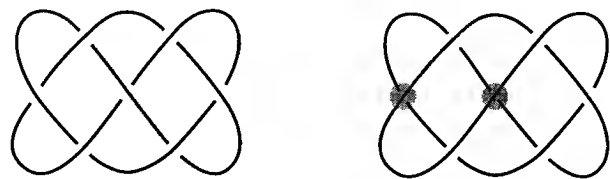
FIGURE 12.43: Unknotting the 7_4 knot.

Besides the topoisomerase II enzyme, there are other enzymes that also perform sophisticated actions on DNA strands. As these enzymes are isolated, knot-theoretic properties can be used to categorize their function.

Once an enzyme is isolated in the laboratory, it is applied to a collection of unknotted cyclic DNA molecules. The various knots that result from its application are then determined, and the action of the enzyme is deduced from the knots. But how are the resulting knots determined? They are so tiny that, although electron microscopes can sometimes be employed to determine the knotting of an individual DNA molecule, there are too many molecules present to use this method effectively.

Fortunately, biologists have developed numerous methods for separating mixtures of molecules on the basis of their chemical and physical properties. For example, if a mixture of variably sized molecules is passed through a porous material, it is not hard to imagine that the smaller molecules will wiggle their way through first. The degree of knottedness in a molecule also influences its ability to traverse such a material. In one such process, called gel electrophoresis, it has been experimentally demonstrated that for small numbers of crossings the speed with which a knot passes through a porous material is proportional to the number of crossings. This technique is used to divide the knot mixture into its constituent knot types. Each type of knot is then imaged using electron microscopy, providing insight into the function of the novel enzyme.

The chemotherapy drug doxorubicin specifically prevents enzymes from unknotting DNA, thereby inhibiting the ability of DNA to replicate (a prerequisite for cell division). Since cancer cells divide very quickly, they are particularly vulnerable to this effect, making the drug an effective cancer-fighter.

Knots in Synthetic Chemistry

In addition to cyclic DNA molecules, much simpler molecules can also include cycles of bonded atoms. A well-known example of a cyclic molecule is benzene, consisting of six carbon and six hydrogen atoms bonded as shown in Figure 12.44.
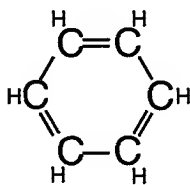
FIGURE 12.44: A diagram of the benzene molecule.

There are many other common cyclic molecules employed in chemistry, and, like benzene, they appear in an unknotted form. However, given a cyclic molecule in an unknotted form, an obvious question is whether or not we can create a knotted version of the molecule as well. In addition, would the knotted version of the molecule have different chemical and physical properties from the unknotted version?

It was relatively recently that chemists first synthesized a knotted molecule. A challenge in so doing was to create a knotted molecule with relatively few atoms (in contrast to DNA, which has millions of atoms and readily forms knots). In 1988, Christina Dietrich-Buchecker and Jean-Pierre Sauvage of

the University of Stausborg became the first chemists to synthesize a knotted molecule. Their process required some creative chemistry, using two copper ions to hold two molecular strands in a particular shape so that when the strands were bonded and the copper ions removed, the resulting molecule was knotted. The molecule consists of 104 carbon, 104 hydrogen, 8 nitrogen, and 14 oxygen atoms. A diagram of the molecule is shown in Figure 12.45; each unlabeled vertex in the diagram represents a carbon atom bonded to either 0, 1, or 2 hydrogen atoms so that each carbon atom has 4 bonds around it (counting each doubled edge as two bonds).
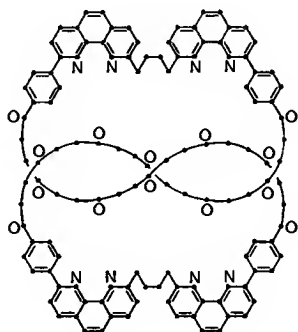
FIGURE 12.45: A diagram of the first synthesized knotted molecule.

Since the first synthesis of a knotted molecule, there has been continued interest in the development and study of techniques for their formation. Once chemists discover simple methods for synthesizing knotted molecules, a new world of chemical products will likely result. Each distinct knot could coincide with a distinct molecule generated by the same set of constituent atoms bonded in the same cyclic order. Given a particular cyclic molecule, the only bound on the number of different knotted molecules that can be made from it is the flexibility of the molecule. The process of distinguishing between knotted molecules and determining their properties will certainly benefit from our understanding of knots and their properties.

The property known as chirality is of significant interest in chemistry, even for molecules that are neither knotted nor cyclic. A molecule that cannot be superimposed on its mirror image is called **chiral**; otherwise, it is called **achiral**. The drug thalidomide is an example of a chiral molecule. Diagrams of (R)-thalidomide and (S)-thalidomide are shown in Figure 12.46. These

(R)-Thalidomide   (S)-Thalidomide
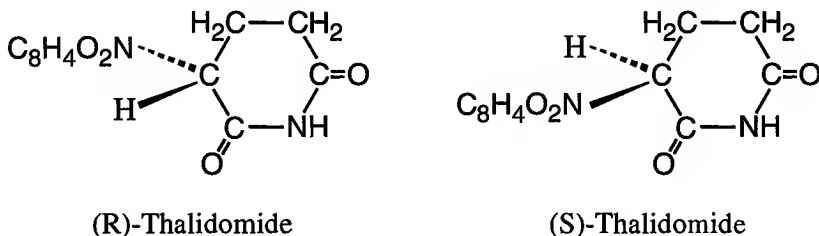
FIGURE 12.46: The molecules (R)-thalidomide and (S)-thalidomide.

two molecules are mirror images of each other by a reflection in the plane of the page. In (R)-thalidomide, on the left in the figure, the solid tapered bond indicates that the H atom is in front of the plane of the page, while the dashed tapered bond indicates that the $C_8H_4O_2N$ group is behind the plane. The situation with (S)-thalidomide is the reverse. If we tried to superimpose (R)-thalidomide onto (S)-thalidomide by rotating the former so that the H atoms and the $C_8H_4O_2N$ groups in each coincide, then the remaining parts of the two molecules would not align properly. Thus, thalidomide is a chiral molecule.

Chemists have long recognized the importance of chirality. A chiral molecule and its mirror image have similar physical properties and similar interactions with achiral molecules, but they can have dramatically different reactions with other chiral molecules. Thalidomide is a classic example. It was prescribed to pregnant women in the 1950s and 1960s to alleviate morning sickness. While (R)-thalidomide is effective in that role, the drug was soon linked to a wave of birth defects. It is now known that (S)-thalidomide, also present in doses of the drug, is responsible for these unfortunate effects. The thalidomide case motivated interest in both the biological chemistry of chiral compounds and the adoption of more stringent drug-safety protocols.

If we take knotted molecules into consideration, then there are important chirality questions that arise just from a knot-theoretic perspective. For instance, if a molecule is represented by a knot that is not equivalent to its mirror image, then the molecule is automatically chiral. Therefore it is of interest to investigate which knots are equivalent to their mirror image.

The mirror image of a knot K is the knot obtained by taking the reflection of K in a plane in $\mathbb{R}^3$. It is straightforward to show that the knot type of the reflection is independent of the plane chosen. If we have a diagram of K, and reflect K in the plane of that diagram, then it follows that the diagram of K can be converted into a diagram of the resulting mirror image by switching all of the crossings. (See Figure 12.47.)
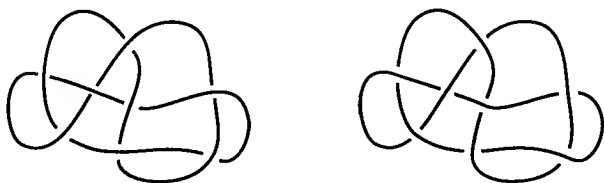
FIGURE 12.47: Diagrams of a knot and its mirror image.

In Figure 12.10 we show a deformation that starts at the figure-eight knot. If you rotate the final stage of the illustration by 180 degrees and deform the knot a little further, it is easy to see that you end up with the mirror image of the knot at the start of the deformation. Therefore the figure-eight knot is equivalent to its mirror image. On the other hand, the trefoil knot ⟨⟩ and its mirror image ⟨⟩ have distinct Kauffman X polynomials and therefore are not equivalent to each other. (See Example 12.10 and Exercise 12.15.)

DEFINITION 12.14. *A knot that is equivalent to its mirror image is said to be* **amphichiral.** *Otherwise, the knot is said to be* **chiral.**

We can use the Kauffman X polynomial to help determine if a knot is chiral. To begin, we have the following theorem that relates the Kauffman X polynomial of a knot to the polynomial of its mirror image:

THEOREM 12.15. *Let K be a knot with Kauffman X polynomial P. If K^* is the mirror image of K, then the Kauffman X polynomial of K^* is obtained by substituting A^{-1} for A in P.*

   Proof. See Exercise 12.23.   ∎

From Theorem 12.15 it follows that if the Kauffman X polynomial of a knot changes under the substitution of A^{-1} for A, then the knot and its mirror image are not equivalent, and therefore the knot is chiral. Polynomials that do not change under a substitution of A^{-1} for A are palindromic, a property of polynomials that is defined as follows:

DEFINITION 12.16. *A polynomial*

$$a_{-n}A^{-n} + a_{-n+1}A^{-n+1} + \ldots + a_{n-1}A^{n-1} + a_n A^n$$

is called **palindromic** *if $a_{-j} = a_j$ for all $j = 1, \ldots, n$.*

Thus a polynomial is palindromic if its coefficients read the same forwards as backwards. It is straightforward to see that a polynomial

$$a_{-n}A^{-n} + a_{-n+1}A^{-n+1} + \ldots + a_{n-1}A^{n-1} + a_n A^n$$

is palindromic if and only if it is unchanged under the substitution of A^{-1} for A.
   We now obtain the following corollary of Theorem 12.15.

COROLLARY 12.17. *If the Kauffman X polynomial of a knot K is not palindromic, then K is chiral.*

Corollary 12.17 yields a simple test for the chirality of knots and knotted molecules. In this way, it is possible to use basic ideas from topology and knot theory to distinguish between molecules.
   Building chemical compounds and distinguishing between them are important aspects of synthetic chemistry. Determining the properties of the resulting compounds is also significant. In Section 13.2, we will show how properties of topological graphs can be used to help predict properties of molecules.

Exercises for Section 12.4

12.19. Describe an algorithm for changing crossings in a link projection of an n-component link so that the result is a link projection of the trivial link with n components.

12.20. Determine the unknotting number of the knots appearing in Figure 12.48.
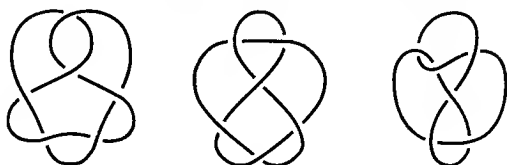
FIGURE 12.48: Determine the unknotting number.

12.21. Show that the unknotting number of each knot in Figure 12.49 is at most 2.
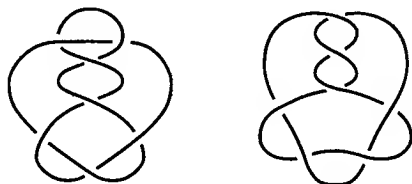
FIGURE 12.49: Show that the unknotting number of each is at most 2.

12.22. Sketch the knot associated with the knotted molecule in Figure 12.45. Then determine the knot in Figure 12.17 that is equivalent to this one by sketching a deformation between the two.

12.23. Prove Theorem 12.15: Let K be a knot with Kauffman X polynomial P. If K^* is the mirror image of K, then the Kauffman X polynomial of K^* is obtained by substituting A^{-1} for A in P. (Hint: First prove the result for the bracket polynomial, then show that it holds for the Kauffman X polynomial.)

12.24. Prove that the knot 5_2, which is the last knot in the top row in Figure 12.17, is chiral.

Chapter 13

Graphs and Topology

In the introduction to this book, we indicated that Euler's solution to the Königsberg bridges problem can be considered the moment of birth of the field of topology. It can equally well be considered the starting point for the field of graph theory. Naturally there is overlap between the two fields, and that overlap is a subdiscipline of each, known appropriately as topological graph theory.

We consider graphs and some of their properties in the first section of this chapter. In Section 13.2, we present an application to the quantitative structure-property relationship in chemistry, where structural aspects of molecules (for example, graph models) are used to predict properties of the molecule. We address important questions and results about embeddings of graphs in Section 13.3. In the last section, we introduce the concepts of crossing number and graph thickness, and we discuss an application in electronic circuit design.

13.1   Graphs

In this section, we introduce some basic properties and results involving graphs. Along the way, we develop appropriate concepts to solve the Königsberg bridges problem via graphs.

In graph theory, a graph G is abstractly defined as a finite set V_G of vertices along with a finite set E_G of unordered pairs $\{v_i, v_j\}$ of vertices. The elements of E_G are called the edges of G.

In topological graph theory, a graph is regarded as a topological space G made up of a finite set of points V_G, called the vertices of G, along with a finite collection E_G of edges joining vertices. Each edge results from taking a closed bounded interval in $\mathbb{R}$ and gluing one endpoint to a vertex v_i and gluing the other endpoint to a vertex v_j. (See Example 3.17.)

There is a straightforward correspondence between the abstract graphs in graph theory and the topological graphs in topological graph theory. Throughout the chapter we do not distinguish between abstract graphs and topological graphs. We use the term graph to refer to both. In cases where the difference matters, it should be clear from the context which meaning is intended.

Let G be a graph. If e is an edge of G with vertices v_i and v_j, then e and the vertices v_i and v_j are said to be **incident to** each other. Also, e is said to **join** the vertices v_i and v_j.

IMPORTANT NOTE: *In a few places in this chapter we construct graphs by removing vertices or edges from a particular graph. In such cases, when we remove an edge from a graph, we are not necessarily removing the vertices incident to the edge unless we explicitly state that we are.*

A graph is a compact Hausdorff space. (See Exercise 13.1.) Therefore each one-point set in a graph is closed. In particular, each vertex is closed as a one-point set. Each edge in a graph is a compact subspace of the graph since it is the image of a compact space under a continuous function (the quotient mapping that glues the edges and vertices together). Since graphs are Hausdorff, it follows that each edge is a closed subset in the graph.

EXAMPLE 13.1. Two important families of graphs are the complete graphs and the complete bipartite graphs. The **complete graph on n vertices** is the graph K_n that has n vertices and a collection of edges such that each pair of distinct vertices is joined by a single edge. (See Figure 13.1.)

The complete graph on n vertices models the handshake problem, which asks, "If n people shake each others' hands, how many handshakes result?"
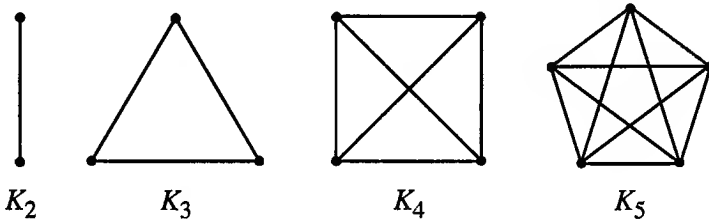
K_2 K_3 K_4 K_5

FIGURE 13.1: The complete graphs K_2, K_3, K_4. and K_5.

The **complete bipartite graph** $K_{m,n}$ is a graph having $m + n$ vertices that can be divided into sets V_m and V_n of m and n vertices, respectively, such that

(i) Each edge joins a vertex in V_m to a vertex in V_n,

(ii) Each pair of vertices $v \in V_m$ and $v' \in V_n$ is joined by a single edge.

Examples of complete bipartite graphs are shown in Figure 13.2.

$K_{1,5}$ $K_{2,3}$ $K_{3,3}$
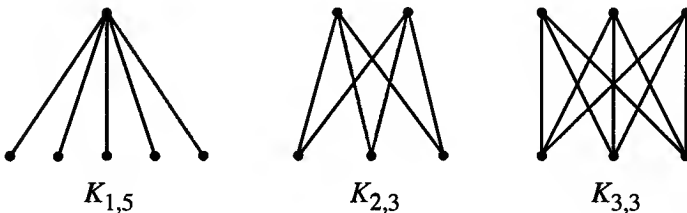
FIGURE 13.2: The complete bipartite graphs $K_{1,5}$, $K_{2,3}$, and $K_{3,3}$.

The graph $K_{3,3}$ models a popular pencil-and-paper puzzle, known as the three-utilities problem, which asks if it is possible (on paper) to connect three houses to three utilities (gas, water, and electricity) without having the lines between the houses and the utilities cross. This amounts to asking if it is possible to embed $K_{3,3}$ in the plane. We discuss this problem further in Section 13.3.

DEFINITION 13.1. *Let G be a graph. For each vertex $v \in G$, define the* ***degree of*** *v to be the number of edges incident to v, counting an edge twice if it joins v to itself.*

We can think of the degree of a vertex v as the number of different approaches to v along edges incident to v.

EXAMPLE 13.2. In Figure 13.3 we display graphs with the degree of each vertex labeled.
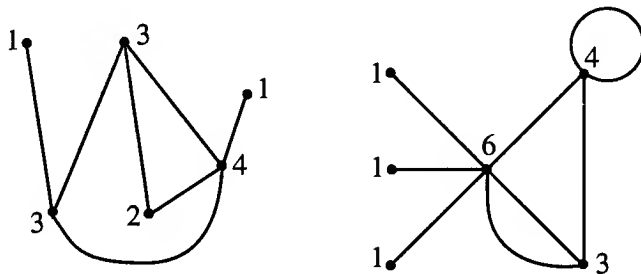
FIGURE 13.3: Degrees of vertices.

EXAMPLE 13.3. In the complete graph on n vertices, K_n, each vertex has degree $n - 1$. In the complete bipartite graph $K_{m,n}$, there are m vertices of degree n that are joined by the graph's edges to n vertices of degree m.

For graphs, equivalence is defined as follows:

DEFINITION 13.2. *Graphs G and G' are* ***isomorphic*** *or* ***graph equivalent*** *if there is a bijective function $f : V_G \to V_{G'}$ between the corresponding sets of vertices, such that for every $v, w \in V_G$, the number of edges in G that join v and w is equal to the number of edges in G' that join $f(v)$ and $f(w)$. Such a function f is called a* ***graph isomorphism.***

Given a graph isomorphism f, it is straightforward to see that f induces a homeomorphism f^* mapping graph G to graph G' and mapping the vertices of G bijectively to the vertices of G'. The converse to this situation also holds. Specifically, we have the following theorem:

THEOREM 13.3. *Let G and G' be graphs with vertex sets V_G and $V_{G'}$, respectively. If there is a homeomorphism $h : G \to G'$ that maps V_G bijectively to $V_{G'}$, then G and G' are isomorphic and the function $h_V : V_G \to V_{G'}$, defined by $h_V(v) = h(v)$, is a graph isomorphism.*

Proof. See Exercise 13.3. ■

If we drop the assumption in Theorem 13.3 that h maps the vertex set of G bijectively to the vertex set of G', then it does not necessarily follow that G and G' are isomorphic. For example, the graphs in Figure 13.4 are homeomorphic, but not isomorphic.
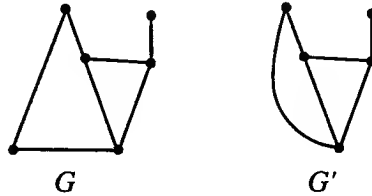
FIGURE 13.4: Graphs G and G' are homeomorphic, but not isomorphic.

Summarizing the preceding definition, theorem, and example, we have two notions of equivalence for graphs G and G':

(i) **Topological Equivalence:** Graphs G and G' are homeomorphic.

(ii) **Graph Equivalence:** Graphs G and G' are homeomorphic by a homeomorphism that maps the vertex set of G bijectively to the vertex set of G'.

In general, we allow graphs to have **parallel edges** (two or more edges that join a pair of vertices) and **loops** (edges that join a vertex to itself). But we are also interested in graphs where neither of these occur:

DEFINITION 13.4. *A graph is said to be **simple** if it has no parallel edges and no loops.*

In Figure 13.5, graphs G_1 and G_2 are simple, but graphs G_3 and G_4 are not.
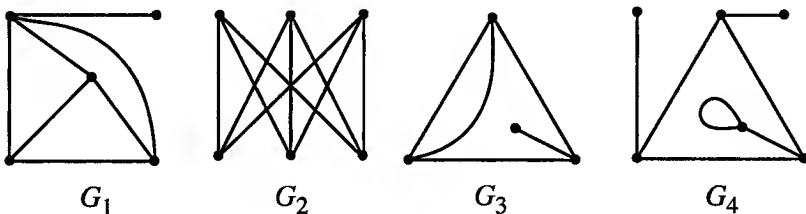
FIGURE 13.5: The first two graphs are simple, but the second two are not.

THEOREM 13.5. *If G is a graph, then there is a simple graph G' that is homeomorphic to G.*

> *Proof.* Let G be a graph. Create a new graph G' as follows. For each edge e in G, joining vertices v and v', add two new distinct vertices, v_1 and v_2, and replace e with three new distinct edges, e_1, e_2, and e_3,

such that e_1 joins v and v_1, e_2 joins v_1 and v_2, and e_3 joins v_2 and v'. (See Figure 13.6.) The resulting graph G' is homeomorphic to G and is simple.   ∎
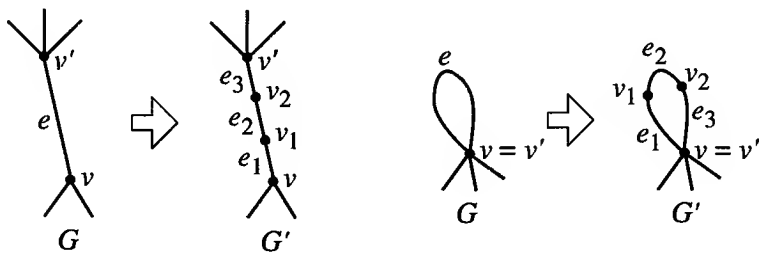
FIGURE 13.6: Adding edges and vertices to make a simple graph.

A walk is the graph–theory equivalent of a path. It is defined as follows:

DEFINITION 13.6. *Let G be a graph and v and v' be vertices of G. A **walk** from v to v' is an alternating sequence of vertices and edges,*

$$v = v_0, e_0, v_1, e_1, \ldots, v_{n-1}, e_{n-1}, v_n = v',$$

such that edge e_i joins vertices v_i and v_{i+1} for each $i = 0, \ldots, n-1$.

The following definition introduces some different types of walks that are of interest in graph theory:

DEFINITION 13.7. *Let G be a graph.*

(i) *A **closed walk** is a walk that begins and ends at the same vertex.*

(ii) *A closed walk in which no edges repeat is called a **circuit**.*

(iii) *A circuit in which no vertices repeat (except the first and last) is called a **cycle**.*

EXAMPLE 13.4. In the graph shown in Figure 13.7, the walk that corresponds to following edges 1–7 in order is a circuit that is not a cycle, and the walk that corresponds to following edges 1–4 in order is a cycle.
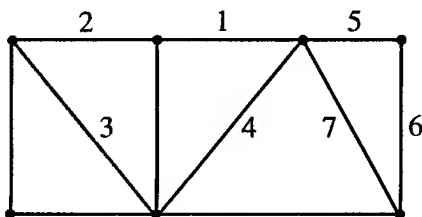
FIGURE 13.7: A circuit (1–7) and a cycle (1–4).

DEFINITION 13.8. *Let G be a graph.*

>   **(i)** *If G has no cycles, then we say that G is **acyclic**.*

>   **(ii)** *If G has at least one cycle, then the **girth** of G is the minimum number of edges in any cycle in G.*

For example, if $m, n \geq 2$, then the girth of the complete bipartite graph $K_{n,m}$ is 4, and if $n \geq 3$, then the girth of the complete graph K_n is 3.

The following theorem provides a condition that guarantees that a graph has at least one cycle:

THEOREM 13.9. *If each vertex in a graph G has a degree of at least 2, then there is a cycle in G.*

>   **Proof.** Pick an edge e_0 of G. Let v_0 and v_1 be the vertices incident to e_0. If $v_0 = v_1$, then we are done, since v_0, e_0, v_1 is then a cycle. Otherwise, let e_1 be an edge incident to v_1 and not equal to e_0. This is possible since each vertex has a degree of at least 2. Now set v_2 equal to the vertex joined to v_1 by e_1. If $v_2 = v_1$ or $v_2 = v_0$, then, as before, we are done. Otherwise, pick an edge e_2 incident to v_2 and not equal to e_1.

>   We continue this process. Since there are only finitely many vertices in G, at some point we must have a walk,

$$v_0, e_0, v_1, e_1, \ldots, v_{m-1}, e_{m-1}, v_m,$$

>   with no repeating edges, with no repeating vertices up through v_{m-1}, and with $v_m = v_k$ for some k between 0 and $m - 1$. Then $v_k, e_k, \ldots, e_{m-1}, v_m$ is a cycle in G.   ∎

DEFINITION 13.10. *Let G be a graph. An **Eulerian walk** in G is a walk that visits each edge in G exactly once. An **Eulerian circuit** in G is a closed Eulerian walk in G. We say that G is **Eulerian** if it contains an Eulerian circuit.*

In Figure 13.8, an Eulerian walk is depicted in the graph on the left (following the edges in numerical order as shown) and an Eulerian circuit is depicted in the graph on the right.
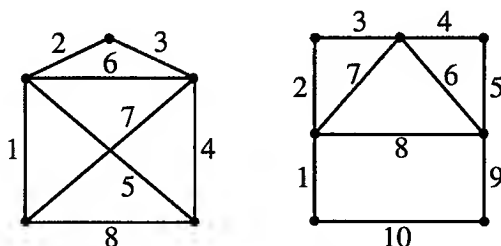
FIGURE 13.8: An Eulerian walk and an Eulerian circuit.

The Königsberg bridges problem, introduced in Section 0.2, asks if there is an Eulerian walk in the associated graph K that models the land regions (vertices) and bridges (edges) in Königsberg. (See Figure 13.9.)
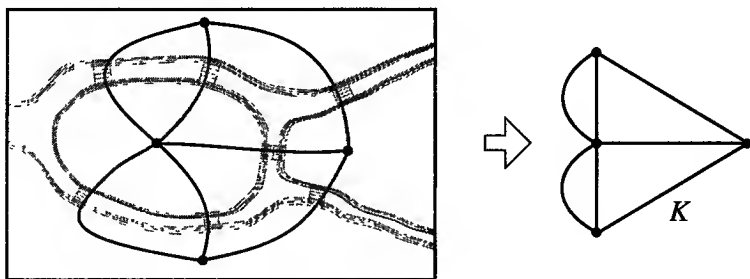
FIGURE 13.9: Modeling Königsberg with graph K.

In his solution to the Königsberg bridges problem (without using the yet-to-be-developed language of graph theory), Leonhard Euler essentially identified necessary and sufficient conditions for a graph to possess an Eulerian walk or circuit, and he proved the necessity. We state the graph theory version of Euler's result in the following theorem:

THEOREM 13.11. *Let G be a graph that is connected.*

   (i) *There is an Eulerian circuit in G if and only if all of the vertices of G have even degree.*

   (ii) *There is an Eulerian walk, but not an Eulerian circuit, in G if and only if G has exactly two vertices with odd degree. Such a walk must begin at one of the vertices with odd degree and must end at the other.*

We prove the necessity implication from Theorem 13.11(i); that is, we prove that if a graph has an Eulerian circuit, then all of the vertices of G have even degree. The converse is the subject of a set of supplementary exercises at the end of the section. The proof for (ii) is similar to that for (i).

Proof of Necessity in (i). Suppose G is a graph and

$$v_0, e_0, v_1, e_1, \ldots, v_{n-1}, e_{n-1}, v_n = v_0$$

is an Eulerian circuit in G. We prove that every vertex of G has even degree.

Let v be a vertex, other than v_0, in the circuit. Every time v is visited on the circuit, it is approached on one edge incident to it and departed from on another. Each of these approaches and departures contributes 1 to the degree of v, and therefore each visit to v counts 2 toward its degree. Since each edge is visited exactly once on a circuit, the degree of v is 2 times the number of visits to v on the circuit. Consequently, the degree of v is even.

For the vertex v_0, the same argument implies that the degree is 1 (for edge e_0) plus 1 (for edge e_{n-1}) plus 2 times the number of visits to v_0 between the start and end of the circuit. Thus the degree of v_0 is also even. ∎

Theorem 13.11 provides us with a solution to the Königsberg bridges problem. Since the Königsberg graph K has four vertices of odd degree, it follows that there is neither an Eulerian circuit nor an Eulerian walk in K. Therefore it is not possible to take a walk through Königsberg that crosses each bridge exactly once.

Exercises for Section 13.1

13.1. Let G be a graph. Prove that G is compact and Hausdorff.

13.2. Determine how many edges there are in the graphs K_n and $K_{n,m}$. Justify your assertion.

13.3. Prove Theorem 13.3: Let G and G' be graphs with vertex sets V_G and $V_{G'}$, respectively. If there is a homeomorphism $h : G \to G'$ that maps V_G bijectively to $V_{G'}$, then G and G' are isomorphic and the function $h_V : V_G \to V_{G'}$, defined by $h_V(v) = h(v)$, is a graph isomorphism.

13.4. Prove that if G is a graph, then every component of G is a graph.

13.5. (a) For which n are the graphs K_n Eulerian? Justify your answer.
   (b) For which m and n are the graphs $K_{m,n}$ Eulerian? Justify your answer.

13.6. Let G be a graph. Prove that G has an even number of vertices having odd degree.

13.7. Draw all of the acyclic graphs (up to graph isomorphism) having eight or fewer edges. Indicate which of these are homeomorphic.

13.8. Let G be an acyclic graph having V vertices and E edges. Prove that $E < V$.

13.9. (a) Prove that if any one of the seven bridges is removed from Königsberg, then it is possible to walk through the city crossing each bridge exactly once.
   (b) Prove that if a new bridge is built between any two distinct land regions in Königsberg, then it is possible to walk through the city crossing each bridge exactly once.
   (c) Prove that you can take any one of the seven bridges in Königsberg and move it to a new location so that after the move has been made, it is possible to walk through the city crossing each bridge exactly once, beginning and ending in the same place.

13.10. For each graph shown in Figure 13.10, use Theorem 13.11 to determine whether the graph has an Eulerian walk, an Eulerian circuit, or neither. If the graph has an Eulerian walk or an Eulerian circuit, sketch it on the graph.
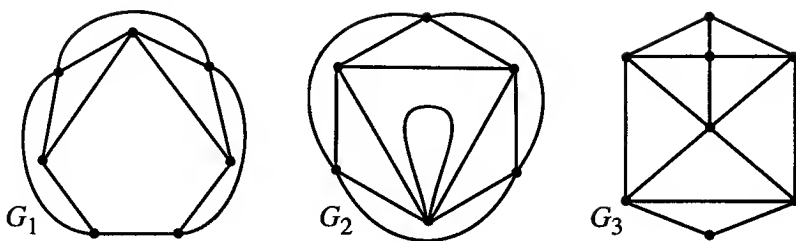
FIGURE 13.10: In each, is there an Eulerian walk or an Eulerian circuit?

Supplementary Exercises: The Existence of an Eulerian Circuit

In these exercises we prove the sufficiency implication in Theorem 13.11(i). That is, we prove that if G is a connected graph and every vertex in G has even degree, then there is an Eulerian circuit in G. The proof proceeds by induction on the number of edges in G.

SE 13.11. Prove the theorem in the case where G has one edge.

Next, assume that G has n edges and that the result is true for graphs having between one and $n - 1$ edges. We prove that there is an Eulerian circuit in G. By Theorem 13.9 there is a cycle

$$v_0, e_0, v_1, e_1, \ldots, v_{j-1}, e_{j-1}, v_j = v_0$$

in G. Each edge in the cycle is visited exactly once, but the cycle need not visit every edge in the graph. Let E^* be the set of edges in the cycle, and let V^* be the set of vertices in the cycle that are incident only to edges in E^*. Set G' equal to the graph obtained from G by removing all of the edges in E^* and all of the vertices in V^*.

SE 13.12. Prove that each vertex in G' has even degree.

Now, G' is not necessarily connected, but each component of it is. Let C be a component of G'. It follows that C is a graph (see Exercise 13.4) and is connected. Furthermore, C has between one and $n - 1$ edges, and each vertex in C has even degree. Therefore, by the induction hypothesis, C has an Eulerian circuit.

SE 13.13. Prove the existence of an Eulerian circuit in G.

Thus, by induction, it follows that if every vertex in a graph G has even degree, then G has an Eulerian circuit.

13.2 Chemical Graph Theory

Graph models are used extensively in chemistry. Typically, the atoms in a molecule are represented by the vertices in a graph, and the bonds are represented by edges. A fundamental goal is to determine the properties of a molecule that can be predicted from the properties of a graph model of it. In this section, we present a particular example where the boiling point of a class of hydrocarbons, known as alkanes, can be accurately predicted using graph models of the molecules.

In chemistry, a descriptor is a modeling tool that is used to predict properties of molecules from quantitative aspects of molecular structure. The collection of descriptors yields what is known as the quantitative structure-property relationship (QSPR). Hundreds of QSPR descriptors have been identified for the analysis of molecules, and they fall into five general categories: constitutional, topological, electrostatic, geometrical, and quantum-chemical. Some of the topological descriptors are derived from properties of graphs modeling molecules. We present one of them here, the Wiener index, named after Harry Wiener who first proposed its use in chemical modeling in 1947. Our presentation of this application is based on the article[Rou].

DEFINITION 13.12. *Let G be a graph.*

    (i) *For vertices v and v' in G, we define the **distance between v and v'** to be the minimum number of edges in a walk from v to v' in G.*

    (ii) *The **Wiener index** of G, denoted W(G), is the sum of the distances between each pair of vertices in G.*

EXAMPLE 13.5. In Figure 13.11, four graphs are shown along with their Wiener indices.

$W(A) = 13$ $W(B) = 27$ $W(C) = 8$ $W(D) = 25$
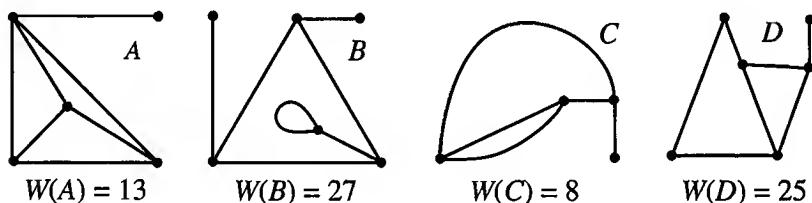
FIGURE 13.11: Wiener indices of graphs.

The graphs C and D shown in Figure 13.11 are topologically equivalent, but their Wiener indices are not equal. Thus the Wiener index is not preserved under topological equivalence. However, as the following theorem indicates, it is preserved under graph isomorphism:

THEOREM 13.13. *If graphs G and G' are isomorphic, then W(G) = W(G').*

    Proof. See Exercise 13.15.   ∎

We apply the Wiener index to the class of molecules known as alkanes. These are straight-chain or branched-chain hydrocarbons in which each bond between carbon atoms is a single, covalent bond. (What we mean by straight-chain and branched-chain should become clear in what follows.)

The simplest alkanes are methane, ethane, and propane. These are all straight-chain hydrocarbons. Representations of each are shown in Figure 13.12.
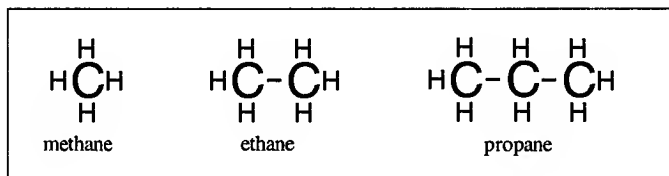
FIGURE 13.12: The alkanes methane, ethane, and propane.

In an alkane, each carbon atom makes a total of four bonds with other carbon and hydrogen atoms. There is a backbone of bonded carbon atoms, and the remaining bonds for each carbon atom are taken up by hydrogen atoms.

From a particular alkane representation, we can form representations that have an additional carbon atom by inserting, between a carbon and hydrogen in the original representation, a new carbon with two hydrogens attached. This is an operation on representations, rather than a chemical process. Thus from the representation for propane, we can form representations of two different alkanes containing four carbons: butane and 2-methylpropane. (See Figure 13.13.) The molecule 2-methylpropane is the simplest branched-chain alkane.
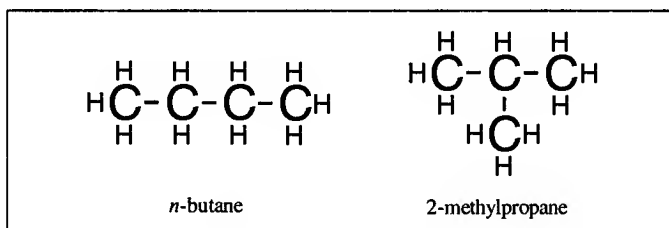
FIGURE 13.13: The two four-carbon alkanes: butane and 2-methylpropane.

From the alkanes with four carbons, we can construct three five-carbon alkanes, as shown in Figure 13.14.
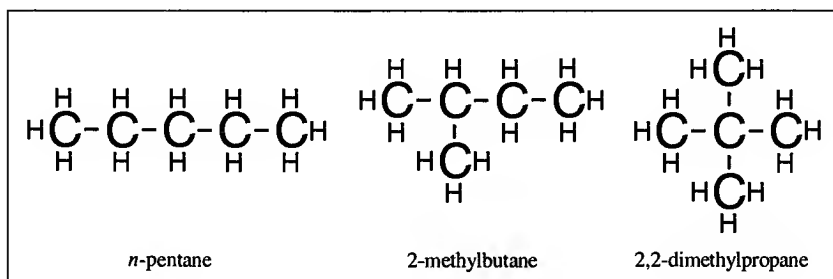
FIGURE 13.14: The three five-carbon alkanes.

We model the alkanes with graphs such that each carbon atom corresponds to a vertex and each bond between carbon atoms corresponds to an edge. Thus the bonded-carbon backbone determines the graph. We ignore the hydrogen atoms and the hydrogen-carbon bonds since they are implicit, given the arrangement of the carbon atoms. The graph models for the alkanes that contain between two and five carbon atoms are shown in Figure 13.15.
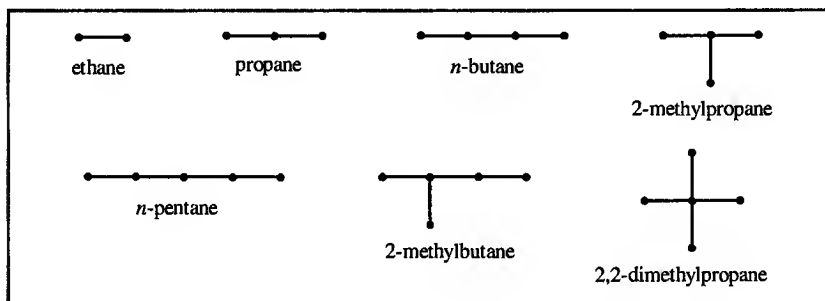
FIGURE 13.15: Graph models of the alkanes with between two and five carbon atoms.

We can construct the graph models for alkanes having six or more carbons by appropriately adding vertices and edges to given alkane graphs. In Figure 13.16, we show the graph models for all of the six- and seven-carbon alkanes.
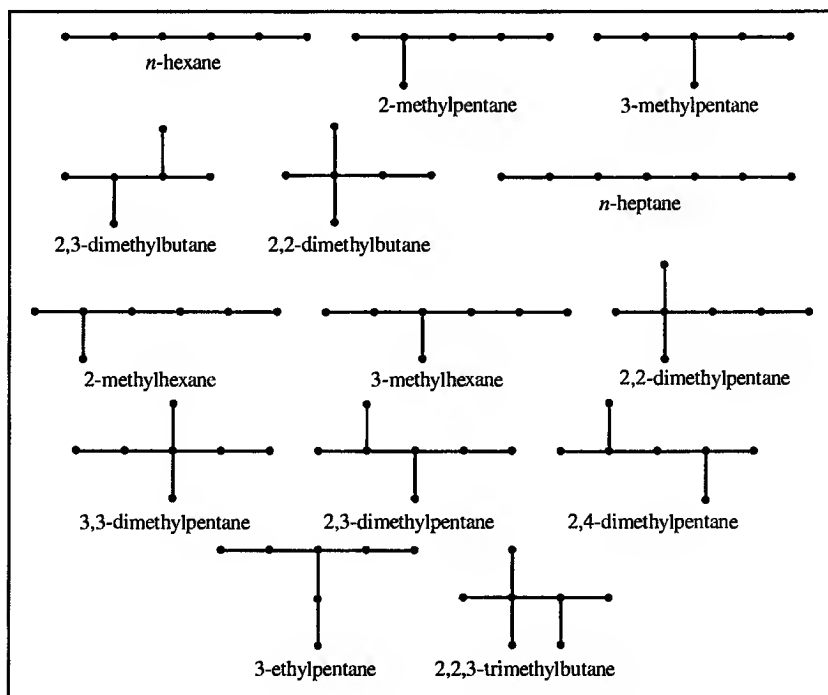
FIGURE 13.16: Graph models of the alkanes with six or seven carbon atoms.

In the table in Figure 13.17, we list all of the alkanes containing between two and seven carbon atoms, and for each we include the boiling point of the molecule (in Kelvin) and the Wiener index of the graph representation of the molecule.

Name	WI	BP (K)			
ethane	1	184	2,3-dimethylbutane	29	331
propane	4	233	2,2-dimethylbutane	28	323
n-butane	10	272	n-heptane	56	371
2-methylpropane	9	261	2-methylhexane	52	363
n-pentane	20	309	3-methylhexane	50	365
2-methylbutane	18	301	2,2-dimethylpentane	46	352
2,2-dimethylpropane	16	283	3,3-dimethylpentane	44	359
n-hexane	35	342	2,3-dimethylpentane	46	363
2-methylpentane	32	333	2,4-dimethylpentane	48	354
3-methylpentane	31	336	3-ethylpentane	48	366
			2,2,3-trimethylbutane	42	354

FIGURE 13.17: Boiling point and Wiener index data for alkanes.

In Figure 13.18, we display a graph of the boiling point and Wiener index data. It is apparent that there is a strong correlation between these two variables. This is the sort of relationship that is beneficial in chemical modeling, a relationship where a quantitative property of a mathematical model of a molecule is closely related to a physical property of the molecule itself.
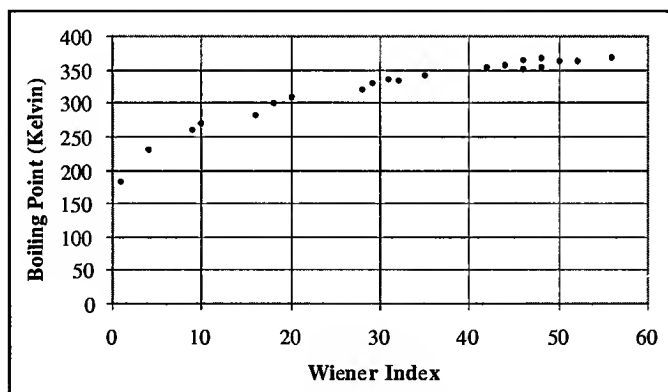
FIGURE 13.18: A graph of the boiling point and Wiener index data.

The boiling point versus Wiener index relationship can be closely approximated by an increasing curve. Fitting a power equation $B = \alpha W^{\beta}$ to the data, we find that the relationship between boiling point (B) and Wiener index (W) for the data is approximated by $B = 181 W^{0.1775}$. (See Figure 13.19.)

Octane is the straight-chain alkane with eight carbon atoms. The Wiener index of its graph representation is 84. Using our power equation, we obtain 397K as an approximation to octane's boiling point. This is in good agreement with the known boiling point of 399K.
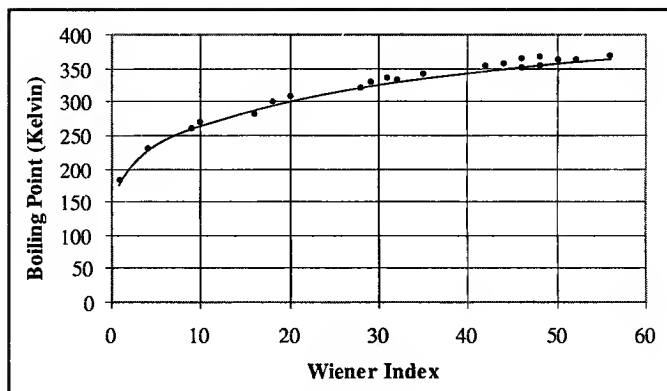
FIGURE 13.19: Fitting a power equation to the data.

Of course, it is not in "predicting" the properties of molecules whose attributes are already known that topological tools such as the Wiener index are useful. It is in predicting the properties and behaviors of molecules yet to be synthesized where such modeling approaches are beneficial. "Will the molecule behave as intended?" and "Will the molecule have minimal toxicity?" are important questions in chemical design. It can be time-consuming and expensive to synthesize molecules and test them in order to answer these questions. The quantitative structure-property relationship and the modeling tools within it provide important information to help anticipate a molecule's characteristics.

Exercises for Section 13.2

13.14. (a) Determine the Wiener index of the complete graph on n vertices, K_n.

   (b) Determine the Wiener index of the complete bipartite graph, $K_{n,m}$.

13.15. (a) **Prove Theorem 13.13:** If graphs G and G' are isomorphic, then $W(G) = W(G')$.

   (b) Show that the converse of Theorem 13.13 does not hold. That is, find graphs G and G' demonstrating that $W(G) = W(G')$ does not necessarily imply that G is isomorphic to G'.

13.16. (a) Determine the Wiener index of the graph $S_n \subset \mathbb{R}$ whose vertices are the integers $1, 2, 3, \ldots, n$, and whose edges are the intervals connecting them, $[1, 2], [2, 3], \ldots, [n - 1, n]$.

   (b) Develop a relationship between the boiling point and number of carbon atoms, n, for the straight-chain alkane A_n whose graph representation is given by the graph S_n in (a).

   (c) Nonane is the straight-chain alkane with nine carbon atoms. Use your relationship from part (b) to approximate the boiling point of nonane. Compare your result with the actual value of 424K.

   (d) Decane is the straight-chain alkane with 10 carbon atoms. Use your relationship from part (b) to approximate the boiling point of decane. Compare your result with the actual value of 447K.

13.17. In the table in Figure 13.20, we list all of the alkanes containing eight carbon atoms, along with their boiling points and associated Wiener indices. Add this data to the data for alkanes containing between two and seven carbon atoms, and use data-analysis software to derive a power equation $B = \alpha W^{\beta}$ approximating the boiling point versus Wiener index relationship for alkanes having between two and eight carbon atoms.

Name	WI	BP (K)			
n-octane	84	399	3,4-dimethylhexane	68	391
2-methylheptane	79.	391	3,3-dimethylhexane	67	385
3-methylheptane	76	392	3-ethylhexane	72	392
4-methylheptane	75	391	2,2,3-trimethylpentane	63	383
2,2-dimethylhexane	71	380	2,3,4-trimethylpentane	65	386
2,3-dimethylhexane	70	389	3-ethyl-2-methylpentane	67	389
2,4-dimethylhexane	71	382	2,2,4-trimethylpentane	66	372
2,5-dimethylhexane	74	382	2,3,3-trimethylpentane	62	388
			3-ethyl-3-methylpentane	64	391

FIGURE 13.20: Boiling point and Wiener index data for the eight-carbon alkanes.

13.3   Graph Embeddings

Graph-embedding questions and results play a major role in topological graph theory. In this section we introduce a few examples of such results, including Kuratowski's Theorem, which provides necessary and sufficient conditions for a graph to be embeddable in the plane.

To begin, we have the following theorem, which indicates that no matter how many vertices and edges a graph has, there is enough room in $\mathbb{R}^3$ to create a copy of the graph:

THEOREM 13.14. *Every graph G can be embedded in $\mathbb{R}^3$.*

Proof. Let G be a graph with n vertices, $v_1, \ldots, v_n$, and m edges, $e_1, \ldots, e_m$. We define an embedding $f : G \to \mathbb{R}^3$. Begin by defining $f(v_i) = (0, 0, i)$, for each $i = 1, \ldots, n$. In this manner, the set of vertices is mapped bijectively to a set of points on the z-axis in $\mathbb{R}^3$.

Let $P_1, \ldots, P_m$ be a set of distinct half planes in $\mathbb{R}^3$, each emanating from the z-axis. Now, if e_k is an edge in G, joining vertices v_i and v_j, then extend f to e_k so that it maps e_k homeomorphically onto the semicircle in P_k running from $(0, 0, i)$ to $(0, 0, j)$. (See Figure 13.21.) The map f, so defined, is an embedding of G in $\mathbb{R}^3$. ∎

Given that all graphs can be embedded in $\mathbb{R}^3$, we next consider embeddings in the plane. Can we still obtain an embedding for every graph? We will see that the answer is no, motivating the following definition:

DEFINITION 13.15. *A graph G is called **planar** if it can be embedded in the plane; otherwise it is called **nonplanar**.*
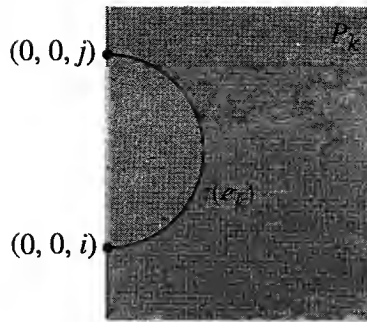
FIGURE 13.21: The edge e_k maps homeomorphically to a semicircle in half plane P_k.

In Section 13.1 we introduced the three-utilities problem, which essentially asks if it is possible to embed the graph $K_{3,3}$ in the plane. After some experimentation, it is apparent that the three-utilities problem does not have a solution, and therefore we conjecture that $K_{3,3}$ is nonplanar. If that is the case, then it follows that no graph $K_{m,n}$ with $m, n \geq 3$ is planar because $K_{3,3}$ can be embedded in each such graph.

On the other hand, we can see from Figure 13.22 that all complete bipartite graphs $K_{1,n}$ and $K_{2,n}$ are planar.
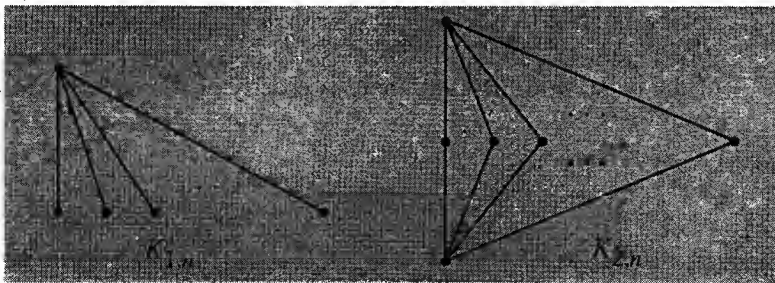
FIGURE 13.22: The complete bipartite graphs $K_{1,n}$ and $K_{2,n}$ are planar.

Now, consider the complete graphs. For $n \leq 4$, we can embed K_n in the plane as illustrated in Figure 13.23. But, if we experiment with K_5, we are led to conjecture that K_5 is nonplanar. And if that is the case, then it follows that none of the complete graphs K_n with $n \geq 5$ are planar since K_5 can be embedded in each of those graphs.

Thus it seems as if $K_{3,3}$, K_5, and every graph containing either $K_{3,3}$ or K_5 as a subspace are nonplanar. The following important topological graph theory result indicates that these are exactly the nonplanar graphs:

THEOREM 13.16. *Kuratowski's Theorem.*   *A graph is planar if and only if it contains no subspace homeomorphic to $K_{3,3}$ and no subspace homeomorphic to K_5.*
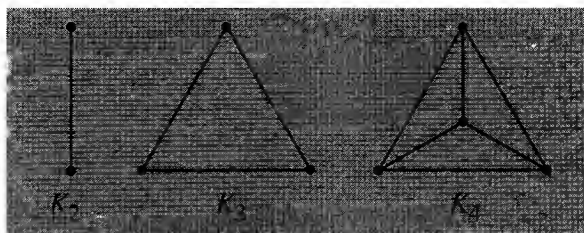
FIGURE 13.23: The complete graphs K_2, K_3, and K_4 are planar.

Kazimierz Kuratowski (1896–1980) was a major contributor to the development of the field of topology in the early twentieth century. In his 1930 paper, "Sur le problème des courbes gauches en topologie," he proved the foregoing theorem that now bears his name. Since 1930 a number of alternative and simplified proofs of the theorem have been published. (See [Tho] and [Mak], for ex:ːiple.)

Here we prove half of Kuratowski's Theorem—that if a graph contains a subspace homeomorphic to either $K_{3,3}$ or K_5 then it is nonplanar. We do not prove the other implication of Kuratowski's Theorem—that if a graph G is nonplanar, then either $K_{3,3}$ or K_5 can be embedded in G; it would take us too far off course to introduce all of the details needed to do so.

If G is an embedded graph in the plane, then the complement of G is made up of a number of components; we call these the **faces** of the embedded graph. (See Figure 13.24.) Furthermore, since G is compact, it is a bounded subset of the plane, and therefore there is exactly one unbounded face of G. (See Exercise 6.19.)
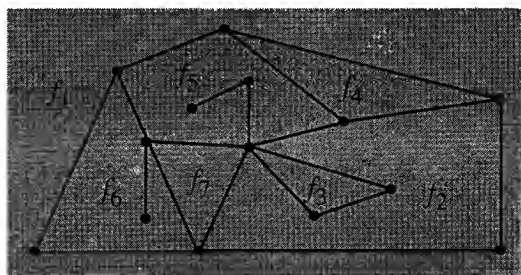
FIGURE 13.24: A graph in the plane separates the plane into faces, f_i.

Consider the planar graphs in Figure 13.24 and Figure 13.25. If we count vertices, edges, and faces, and let the total of each be V, E, and F, respectively, then in all three cases we obtain the relationship $V - E + F = 2$. It is not a coincidence that the result of $V - E + F$ is 2 in each of these examples. This relationship is a consequence of Euler's formula for planar graphs, a result that we establish in Theorem 13.18.

Our proof of Euler's formula is based on the next theorem, which is a general separation and nonseparation theorem for graphs in the plane that builds on the Jordan Curve Theorem (Theorem 11.2) and Theorem 11.10.
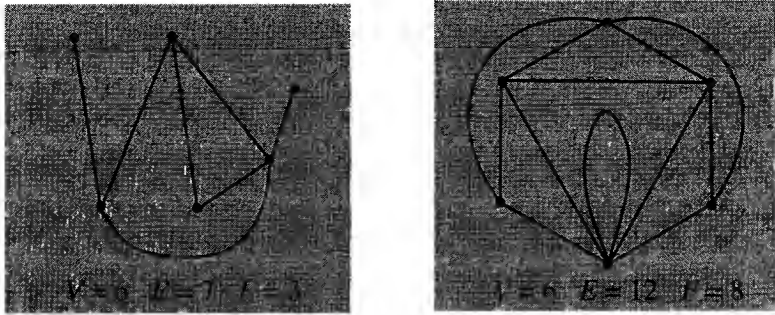
FIGURE 13.25: The relationship V - E + F = 2 holds for both of these graphs.

Although the theorem is intuitively clear, like the Jordan Curve Theorem it is not straightforward to prove. We do not include a proof.

The setup for the theorem is as follows: Let G be a graph in the plane. We are going to add an edge e to G to obtain a new graph G' in the plane. We assume that the vertices of e lie in G, but otherwise e does not intersect G. It follows that the interior of e (relative to G') lies in a face f of G. We are interested in knowing how the addition of e to the graph impacts the face f.

THEOREM 13.17. *Let G, e, f, and G' be as given.*

    (i) *If the vertices of e lie in separate components of G, then $f - e$ is a face f' of G', and e lies in the boundary of f'.*

    (ii) *If the vertices of e lie in the same component of G, then $f - e$ is made up of two faces of G', f' and f'', and e lies in the boundary of both f' and f''.*

Thus, in the first case in Theorem 13.17 the new edge e does not separate the face f of G, but in the second case it does. Note that in the second case it is possible that the vertices coincide. We illustrate the two cases from Theorem 13.17 in Figure 13.26, and we have drawn the edges as curves, rather than line segments, to emphasize the fact that this result is about general embeddings of graphs in the plane, not just those whose image is made up of line segments.
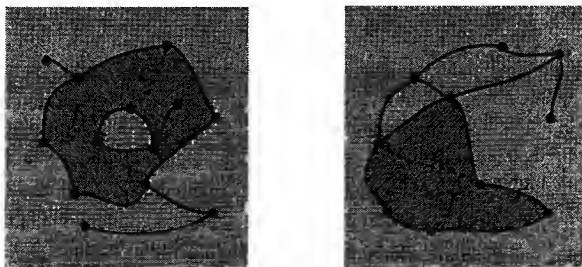
FIGURE 13.26: In the first case the new edge e does not separate f, but in the second case it does.

The relationship $V - E + F = 2$ holds for planar graphs that are connected. Euler's formula for planar graphs, which we prove next, holds for all graphs in the plane. Note that if a graph is not connected, then it is a union of finitely many components, each of which is a connected graph. (See Exercise 13.4.)

THEOREM 13.18. *Euler's Formula for Planar Graphs.* *Let G be a nonempty graph in the plane having C components. If there are V vertices, E edges, and F faces associated with G, then $V - E + F = C + 1$.*

> ***Proof.*** We prove the result by induction on the number of edges E. If $E = 0$, then G consists of V vertices and no edges. Thus G has V components; that is, $C = V$. Since the plane with a finite set of points removed is connected (see Exercise 6.43), G has a single face associated to it. Thus, $F = 1$. We have $V - E + F = V + 1 = C + 1$. Therefore Euler's formula holds in the case where $E = 0$.
>
> Now assume that Euler's formula holds for graphs G that lie in the plane and have $E - 1$ edges. Let G' be a graph that lies in the plane and has $E \geq 1$ edges. Assume that V, F, and C are the number of vertices, faces, and components, respectively, associated with G'. We consider two cases: either G' is acyclic or it is not.
>
> First assume that G' is acyclic. Pick a component H of G' containing at least one edge. The component H is an acyclic graph, and therefore it follows from Theorem 13.9 that H has a vertex of degree 1. Pick such a vertex v and let e be the edge incident to it. Remove e from G' to obtain a graph G. The vertex v is a single-point component of G, and G has one component more than G' has. Furthermore, G has V vertices and $E - 1$ edges. If we add e back, obtaining G' from G, then Theorem 13.17 implies that the number of faces is the same for G' and G. Therefore G has F faces. Applying the inductive hypothesis to G, we have $V - (E - 1) + F = (C + 1) + 1$. Thus, in this case we obtain $V - E + F = C + 1$, as desired.
>
> Now assume that G' has a cycle. Let e be an edge in a cycle in G'. Remove e from G' to obtain a graph G. Since e lies in a cycle, the removal of e from G' does not change the number of components of the graph, and the vertices of e lie in the same component of G. Furthermore, G has V vertices and $E - 1$ edges. If we add the edge e back, obtaining G' from G, then Theorem 13.17 implies that the number of faces of G' is one more than the number of faces of G. Thus G has $F - 1$ faces. Applying the inductive hypothesis to G, it follows that $V - (E - 1) + F - 1 = C + 1$; therefore $V - E + F = C + 1$.
>
> In both cases we obtain $V - E + F = C + 1$, and therefore Euler's formula holds for graphs that lie in the plane and have E edges, assuming that it holds for graphs that lie in the plane and have $E - 1$ edges. Thus, by induction, the theorem follows. ∎

A nice consequence of Euler's formula for planar graphs is Euler's formula for polyhedra. If we count the number of faces F, edges E, and vertices V of a polyhedron, as in Figure 13.27, then we obtain $V - E + F = 2$. We can see why this holds by removing a point from a face in the polyhedron and then mapping the resulting space homeomorphically to the plane. The vertices and edges of the polyhedron map to a connected graph in the plane, and the faces of the polyhedron map to the faces of the graph. Therefore, by Theorem 13.18 it follows that $V - E + F = 2$.

$V = 6, E = 12, F = 8$ $V = 8, E = 12, F = 6$

$V = 20, E = 30, F = 12$

$V = 6, E = 9, F = 5$ $V = 9, E = 14, F = 7$
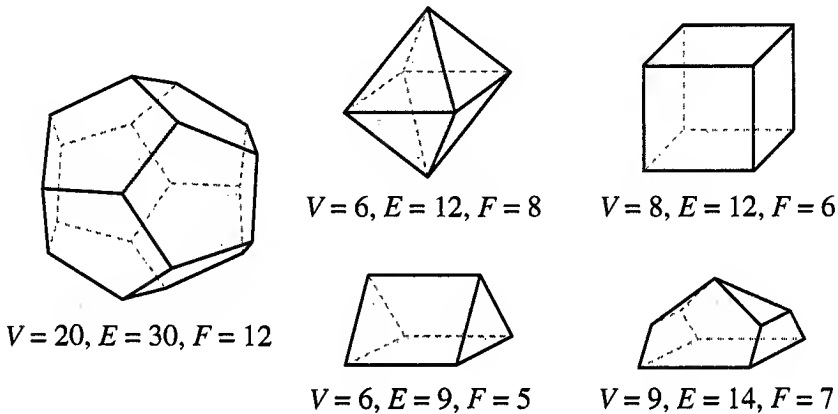
FIGURE 13.27: The relationship $V - E + F = 2$ holds for polyhedra.

The following lemmas are subsequently employed to derive a useful bound (Theorem 13.21) on the number of edges in a planar graph:

LEMMA 13.19. *Let G be a graph in the plane.*

(i) *If e is an edge that does not lie in a cycle in G, then e lies in the boundary of a single face of G.*

(ii) *If e is an edge that lies in a cycle in G, then e lies in the boundary of two faces of G.*

Proof. See Exercise 13.22. ∎

LEMMA 13.20. *Let G be a graph in the plane. If G has more than one face, then the boundary of every face of G contains a cycle.*

Lemma 13.20 does not imply that the boundary of each face is a cycle, but just that each face has a cycle in its boundary. Examining Figure 13.24, we see examples where the boundary of a face is not a cycle. In particular, in the boundary of face f_5 and in the boundary of face f_6 there are edges that are not part of any cycle in the graph, and in the boundary of face f_2 there are two cycles.

Proof. Assume that G has more than one face, and let f be a face of G. Let H be the graph made up of all of the edges and vertices of G that lie in the boundary of f. We claim that there is a cycle in H, and therefore there is a cycle of edges in G that lies in the boundary of f.

First, we note that f is a face of H, and H has more than one face. Let p be a path from a point in f to a point in another face of H. If α is the last point on p in $\mathrm{Cl}(f)$, then α lies on an edge e of H, α lies in the boundary of f, and α lies in the boundary of at least one other face of H. It follows that e is in the boundary of more than one face of H, and therefore Lemma 13.19 implies that e lies in a cycle in H. Thus H contains a cycle, as we wished to show.   ∎

The following theorem provides us with a bound on the number of edges in a planar graph, expressed in relation to the number of vertices in the graph and the girth of the graph. Recall that the girth of a graph is the minimum number of edges in any cycle in the graph.

THEOREM 13.21. *Let G be a planar graph having V vertices, E edges, and at least one cycle. If g is the girth of G, and $g \geq 3$, then*

$$E \leq \frac{g(V-2)}{g-2}.$$

Proof. Since G is planar, it can be embedded in the plane. We may assume that G itself lies in the plane. Each face associated with G has a cycle in its boundary (by Lemma 13.20), and each cycle consists of at least g edges; therefore we can think of each face as contributing at least g edges to G. But since each of these gF edges lies in a cycle, Lemma 13.19 implies that each such edge arises from two different faces. So the number of edges in G is at least $gF/2$. Hence $gF/2 \leq E$, and therefore $F \leq 2E/g$.

Substituting for F in Euler's formula, we obtain

$$2 \leq C+1 = V - E + F \leq V - E + \frac{2E}{g}.$$

Therefore,

$$E - \frac{2E}{g} \leq V - 2,$$

implying that

$$gE - 2E \leq g(V-2).$$

Since $g \geq 3$, it follows that

$$E \leq \frac{g(V-2)}{g-2}.$$

    ∎

COROLLARY 13.22. *The graphs $K_{3,3}$ and K_5 are nonplanar.*

Proof. For $K_{3,3}$ we have $V = 6$, $E = 9$, and $g = 4$. If we could embed $K_{3,3}$ in the plane, then by Theorem 13.21 we would have $E \leq 2V - 4$. But E and V do not satisfy this inequality; therefore $K_{3,3}$ is nonplanar.

Now consider K_5. Here $V = 5$, $E = 10$, and $g = 3$. Since E and V do not satisfy the inequality $E \leq 3V - 6$, it follows that K_5 is nonplanar as well. ∎

Corollary 13.22 yields half of Kuratowski's Theorem. The graphs $K_{3,3}$ and K_5 are nonplanar, and therefore if G is a graph containing either of these graphs as a subspace, then G is nonplanar as well.

EXAMPLE 13.6. The Harary graphs $H_{4,6}$ and $H_{4,7}$ are shown in Figure 13.28. Are either of them planar?
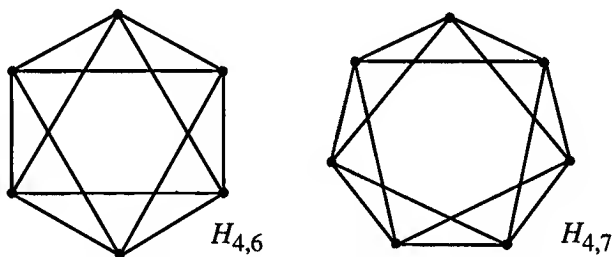
FIGURE 13.28: The Harary graphs $H_{4,6}$ and $H_{4,7}$.

For $H_{4,6}$, the answer is yes. In Figure 13.29 we show an embedding of $H_{4,6}$ in the plane, obtained simply by rearranging three of the edges in the representation of $H_{4,6}$ shown in Figure 13.28.
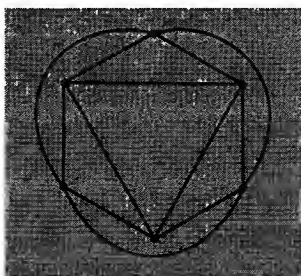
FIGURE 13.29: An embedding of $H_{4,6}$ in the plane.

For $H_{4,7}$, the answer is no. In Figure 13.30 we illustrate an embedding of $K_{3,3}$ in $H_{4,7}$. By Kuratowski's Theorem, it follows that $H_{4,7}$ is nonplanar.
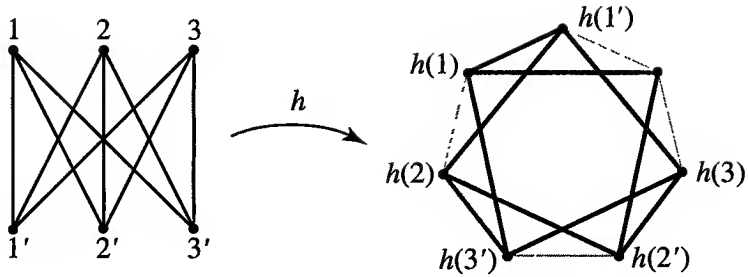
FIGURE 13.30: The bipartite graph $K_{3,3}$ embeds in $H_{4,7}$.

Exercises for Section 13.3

13.18. (a) Provide illustrations demonstrating that K_5 and $K_{3,3}$ can be embedded in a Möbius band.

(b) Provide illustrations demonstrating that K_5 and $K_{3,3}$ can be embedded in a torus.

13.19. Let G be an acyclic graph. Prove that G is planar.

13.20. Use Theorem 13.14 and Corollary 13.22 to prove that $\mathbb{R}^2$ and $\mathbb{R}^3$ are not homeomorphic.

13.21. Although $K_{3,3}$ is a simple graph that has six vertices and nine edges and is not planar, there exist other simple graphs that have six vertices and nine edges and are planar. Find one, and verify that (unlike $K_{3,3}$) it satisfies the inequality in Theorem 13.21.

13.22. Prove Lemma 13.19: Let G be a graph in the plane.

(a) If e is an edge that does not lie in a cycle in G, then e lies in the boundary of a single face of G.

(b) If e is an edge that lies in a cycle in G, then e lies in the boundary of two faces of G.

13.23. (a) Use Theorem 13.21 to prove that the Heawood graph H, shown in Figure 13.31, is nonplanar.

(b) Since the Heawood graph is nonplanar, Kuratowski's Theorem implies that it must contain either a subspace homeomorphic to $K_{3,3}$ or a subspace homeomorphic to K_5. Sketch such a subspace in a diagram of the Heawood graph.
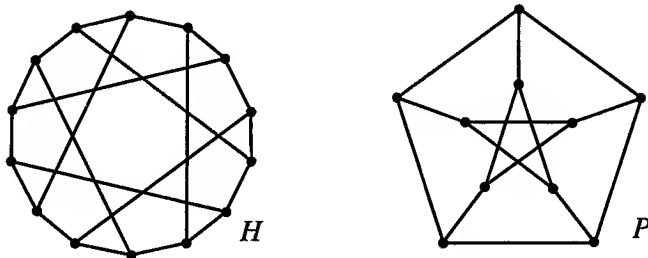
FIGURE 13.31: The Heawood graph H, and the Petersen graph P.

13.24. **(a)** Use Theorem 13.21 to prove that the Petersen graph P, shown in Figure 13.31, is nonplanar.

(b) Since the Petersen graph cannot be embedded in the plane, Kuratowski's Theorem implies that it must contain either a subspace homeomorphic to $K_{3,3}$ or a subspace homeomorphic to K_5. Sketch such a subspace in a diagram of the Petersen graph.

13.25. Construct a graph G as follows: Place a vertex on each square in a 4×4 chessboard. Join two vertices with an edge if there is a knight move from the square associated with one vertex to the square associated with the other. (A knight move goes either two units horizontally and one unit vertically, or two units vertically and one unit horizontally.) Sketch the graph G, and either illustrate an embedding of G in the plane or prove that G is nonplanar.

13.4   *Crossing Number and Thickness*

In this section we introduce crossing number and thickness, two commonly studied measures of a graph's failure to be planar. They are central topics in topological graph theory, and, as we show in an example, they are important considerations for the design of electronic circuits.

Essentially, the crossing number of a graph is the minimum number of crossings needed in a drawing of the graph on paper. To make this formal, we need to be clear about what is meant by "drawing" and by "crossing."

First, a **drawing** of a graph G is a continuous function $f : G \to \mathbb{R}^2$. Of course, this notion of a drawing is too broad because it allows for the inclusion of fairly undesirable images of the graph in the plane. We want images that are as representative of the graph as possible, with as few crossings as possible. Thus, we restrict further:

DEFINITION 13.23. *A drawing* $f : G \to \mathbb{R}^2$ *of a graph is called a good drawing if it satisfies the following conditions:*

    (i) No point in the image of f corresponds to more than two points of the graph;

    *(ii) There are only finitely many points in the image of f that correspond to two points of the graph. These are called **crossing points** of the drawing;*

    (iii) No crossing point of the drawing corresponds to a vertex of the graph.

THEOREM 13.24. *Every graph G has a good drawing.*

Proof. Suppose we have a graph G. As in the proof of Theorem 13.14, we can embed G in $\mathbb{R}^3$ so that the vertices of G map to points on the z-axis and the edges of G map to semicircles in half planes emanating from the z-axis. If we compose the embedding into $\mathbb{R}^3$ with a projection onto a plane P in $\mathbb{R}^3$, we then have a drawing of G, but not necessarily a good drawing. However, if we choose P so that it is parallel to the z-axis and

is not perpendicular to any of the half planes containing the images of the edges of G, then the resulting drawing will satisfy, at least, requirements (ii) and (iii) to be a good drawing. Then, if there are points in the resulting drawing that correspond to more than two points in the graph, we can deform the corresponding semicircles slightly so that requirement (i) for a good drawing is satisfied as well.   ∎

Since all graphs have good drawings, we have the following definition.

DEFINITION 13.25. *Given a graph G, the **crossing number** of G, denoted $v(G)$, is the minimum number of crossing points in any good drawing of G.*

The following theorem describes two essential facts concerning the crossing number:

THEOREM 13.26.

  (i) *Crossing number is a topological invariant of graphs; that is, if graphs G_1 and G_2 are homeomorphic, then $v(G_1) = v(G_2)$.*

  (ii) *A graph can be embedded in the plane if and only if its crossing number is 0.*

Proof. The second part of the theorem is immediate. Consider the first part. Suppose that we have a good drawing $d_1 : G_1 \to \mathbb{R}^2$. We claim that there is a homeomorphism $h : G_2 \to G_1$ such that the function d_2, given by $d_2 = d_1 \circ h$, is a good drawing of G_2. Thus, for every good drawing d_1 of G_1 there is a corresponding good drawing d_2 of G_2 such that the images of d_1 and d_2 are identical. It follows that the crossing numbers of G_1 and G_2 are equal.

To complete the proof, we need to establish the existence of the homeomorphism $h : G_2 \to G_1$. Given $h^* : G_2 \to G_1$, a homeomorphism, the only way that $d_1 \circ h^*$ could fail to yield a good drawing is if h^* maps one or more degree-2 vertices of G_2 to points of G_1 that correspond to crossing points of the drawing d_1. In such an instance, the homeomorphism h^* can be adjusted slightly to a homeomorphism h mapping all degree-2 vertices of G' to points of G that do not correspond to crossing points. (See Exercise 13.26.)   ∎

EXAMPLE 13.7. As we saw in the last section, each of the graphs K_5, $K_{3,3}$, and $H_{4,7}$ is nonplanar, and therefore by Theorem 13.26 the crossing number of each is nonzero. In Figure 13.32 we show good drawings of each of these graphs, and in each case there is only one crossing. Thus the crossing number of each of these graphs is 1.
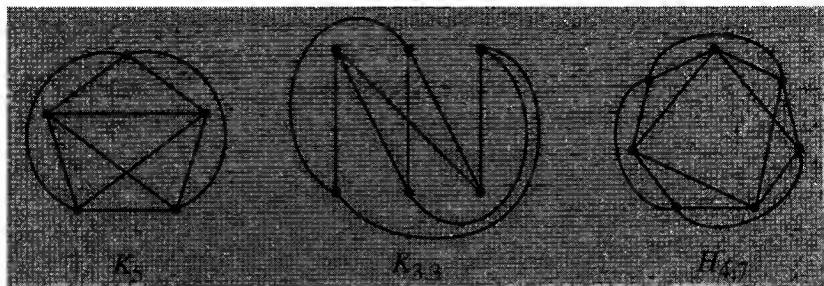
FIGURE 13.32: The graphs K_5, $K_{3,3}$, and $H_{4,7}$ have crossing number 1.

EXAMPLE 13.8. Consider the graph $K_{3,4}$ shown in Figure 13.33. Since $K_{3,3}$ embeds in $K_{3,4}$, the crossing number of $K_{3,4}$ is at least 1. The good drawing of $K_{3,4}$ in Figure 13.33 has two crossings; therefore the crossing number of $K_{3,4}$ is at most 2. Could the crossing number be 1?
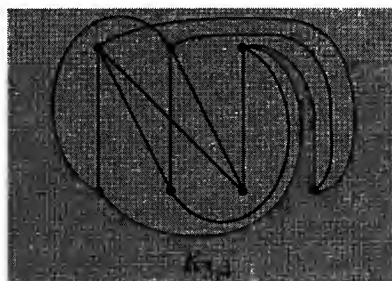
FIGURE 13.33: The graph $K_{3,4}$ has crossing number 2.

We argue by contradiction that $K_{3,4}$ cannot have crossing number 1. Suppose we have a good drawing of $K_{3,4}$ with only one crossing. Then if we remove one of the edges involved in the crossing, the resulting graph is planar (since we are eliminating the only crossing). However, if any one of the edges is removed from $K_{3,4}$, the result is still a graph in which $K_{3,3}$ can be embedded, and therefore $K_{3,4}$ with an edge removed is nonplanar. This is a contradiction, implying that there does not exist a good drawing of $K_{3,4}$ with only one crossing. Thus the crossing number of $K_{3,4}$ is 2.

The problem of determining the crossing number for the complete bipartite graphs $K_{m,n}$ is known as Turan's brickyard problem, named after Paul Turan (1910–1976), a major contributor to graph theory. Turan encountered this problem in 1944, while working in a brick factory in a war labor camp near Budapest, Hungary. In [Tur], a welcome note in the first issue of the *Journal of Graph Theory*, he described the situation that led to his investigation of crossing numbers:

There were some kilns where the bricks were made and some open storage yards where the bricks were stored. All the kilns were connected by rail with all the storage yards. The bricks were carried on small wheeled trucks to the storage yards. All we had to do was put the bricks on the trucks at the kilns, push the trucks to the storage yards, and unload them there. We had a reasonable piece rate for the trucks, and the work itself was not difficult; the trouble was only the crossings. The trucks generally jumped the rails there, and the bricks fell out of them; in short, this caused a lot of trouble and loss of time which was rather precious to all of us (for reasons not to be discussed here). We were all sweating and cursing at such occasions, I too; but *nolens-volens* the idea occurred to me that this loss of time could have been minimized if the number of crossings of the rails had been minimized. But what is the minimum number of crossings? I realized after several days that the actual situation could have been improved, but the exact solution of the general problem with m kilns and n storage yards seemed to be very difficult and again I postponed my study of it to times when fears for my family would end.

Although it has been investigated by many mathematicians since it was first posed by Turan, the problem of determining a general formula for the crossing number for $K_{n,m}$ is still unsolved.

While a graph with a crossing number of at least 1 cannot be embedded in the plane, the next theorem indicates that it can be embedded in a surface obrained by taking a connected sum of tori. (See Section 3.4.)

THEOREM 13.27. *If graph G has crossing number n, then G can be embedded in the connected sum of n tori.*

Proof. Take a good drawing of G with n crossings. Using the inverse of stereographic projection, we can map the good drawing from the plane to the sphere. In a neighborhood of each crossing in the resulting image in the sphere, put a tube (as illustrated in Figure 13.34), with one of the crossing strands running over the tube and the other crossing strand running through the tunnel in the surface that results from the addition of the tube. In this way the crossings have been removed, and the result is an embedding of G in a sphere with n tubes, a surface homeomorphic to a connected sum of n tori.   ∎
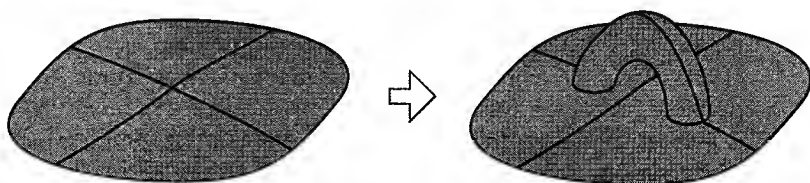
FIGURE 13.34 Removing a crossing by adding a tube.

Another measure of the failure of a graph to be planar is thickness, defined as follows:

DEFINITION 13.28. *The* ***thickness*** *of a graph G, denoted $\theta(G)$, is the smallest n such that G can be expressed as a union of n planar graphs.*

It follows directly from the definition of thickness that a graph is planar if and only if its thickness is 1.

EXAMPLE 13.9. The graphs $K_{3,3}$ and K_5 have thickness 2. Why? Since these graphs are not planar, they each have a thickness of at least 2. Let K' be the graph consisting of one edge from $K_{3,3}$, along with the vertices incident to that edge, and let K'' be the graph consisting of the rest of the edges in $K_{3,3}$, along with all of the vertices in $K_{3,3}$. Each of the graphs K' and K'' is planar, and $K_{3,3}$ is the union of these two graphs; therefore $\theta(K_{3,3}) = 2$. Similarly, it follows that $\theta(K_5) = 2$.

It is evident that thickness is invariant under graph isomorphism. However, thickness is not invariant under homeomorphism; that conclusion follows from the next two theorems.

THEOREM 13.29. *Let G be a nonplanar graph. There exists a graph G' that has thickness 2 and is homeomorphic to G.*

> **Proof.** Consider the graph G' that is constructed from G by trisecting every edge in G, as in the proof of Theorem 13.5. (See Figure 13.6.) The graphs G and G' are homeomorphic.
>
> Now, let G_1 be the graph made up of all of the middle-third edges in G', along with all of the vertices incident to those edges. Let G_2 be the remainder of the edges in G', along with all of the vertices incident to those edges. Each of G_1 and G_2 is planar (see Exercise 13.30), and it follows that G' has thickness 2.   ◘

By Theorem 13.29, every nonplanar graph is homeomorphic to a graph with thickness 2. Does every nonplanar graph have thickness 2 itself? The answer is no; one consequence of the next result is that there are graphs with arbitrarily large thickness. (See Corollary 13.31 and Exercise 13.31.)

THEOREM 13.30. *Let G be a graph having E edges, V vertices, and girth $g \geq 3$. Then*

$$\theta(G) \geq \frac{E(g-2)}{g(V-2)}.$$

Proof. Suppose that the thickness of G is $\theta(G)$. Then G can be expressed as the union of $\theta(G)$ planar graphs G_i. We may assume that no two of the graphs G_i have an edge in common; otherwise we can remove a common edge from all but one of the graphs containing that edge, and we would still have G expressed as a union of $\theta(G)$ planar graphs. Let graph G_i have E_i edges and V_i vertices.

Suppose that graph G_i has a cycle. If the girth of G_i is g_i, then Theorem 13.21 implies that

$$E_i \leq \frac{g_i(V_i - 2)}{g_i - 2} \leq \frac{g(V - 2)}{g - 2}.$$

The second inequality holds since $V_i \leq V$, $g_i \geq g$, and the value of $\frac{g}{g-2}$ decreases as g increases.

Now suppose that G_i is acyclic. We claim that $E_i \leq \frac{g(V-2)}{g-2}$ in this case as well. To begin, we note that $E_i < V_i$ since G_i is acyclic. (See Exercise 13.8.) Also, $g \leq V$ generally holds, and therefore $\frac{V}{V-2} \leq \frac{g}{g-2}$, which implies that $V \leq \frac{g(V-2)}{g-2}$. Thus

$$E_i \leq V_i \leq V \leq \frac{g(V - 2)}{g - 2},$$

as we wished to show.

Now, if we sum the inequalities $E_i \leq \frac{g(V-2)}{g-2}$ over all G_i, then we obtain

$$E \leq \theta(G)\frac{g(V - 2)}{g - 2},$$

and from that the desired inequality follows. ∎

From Theorem 13.30 we obtain the following corollary regarding the thickness of the complete graphs and the complete bipartite graphs.

COROLLARY 13.31. *For K_n and $K_{n,m}$, thickness is bounded from below as follows:*

(i) If $n \geq 3$, then $\theta(K_n) \geq \frac{n(n-1)}{6(n-2)}$.

(ii) If $n, m \in \mathbb{Z}_+$ and $n + m \geq 3$, then $\theta(K_{n,m}) \geq \frac{nm}{2(n+m-2)}$.

Proof. See Exercise 13.29. ∎

EXAMPLE 13.10. By Corollary 13.31, the thickness of the complete bipartite graph $K_{7,7}$ is bounded from below by

$$\frac{(7)(7)}{2(7+7-2)} = \frac{49}{24}.$$

Therefore $\theta(K_{7,7}) \geq 3$. In Figure 13.35 we present three planar graphs whose union is $K_{7,7}$. In the figure, we view $K_{7,7}$ as the graph whose edges join each of the seven lettered vertices $(A-G)$ to each of the other seven vertices. It follows that $\theta(K_{7,7}) = 3$.
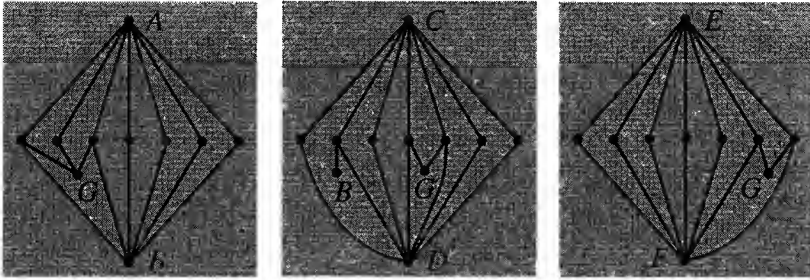
FIGURE 13.35: The bipartite graph $K_{7,7}$ is the union of the three planar graphs shown.

EXAMPLE 13.11. An Application to Electronic Circuit Design. Printed circuit boards are one of the significant inventions that led to the rapid advance in electronic technology in the latter half of the twentieth century. A printed circuit board is a board (or collection of boards) holding electronic components that are connected to each other by conducting pathways, called etches or traces. Circuit boards are found in many electronic devices, from personal computers to DVD players to portable music players. They can hold up to thousands of components including resistors, capacitors, inductors, transistors, and integrated circuit chips (which themselves are miniature circuits).

Each component on a printed circuit board has a number of wire leads that connect it to the conducting pathways. Components can have as few as one or two leads or, in the case of integrated circuit chips, as many as hundreds of them. The components are mounted on the printed circuit board and their leads are connected by conducting material, usually strips of copper that are etched into the board.

Figure 13.36 illustrates a circuit diagram for a simple AM radio receiver. Usually, the characteristics of each component (for example, the voltage of a power supply) would be labeled in a such a diagram, but it is only the topological structure of the circuit that concerns us here. Notice that there is one place, toward the lower left in the diagram, where one conducting pathway passes over another. Is it possible to avoid having such an overpass by rearranging the circuit? We answer that question by considering a graph model.
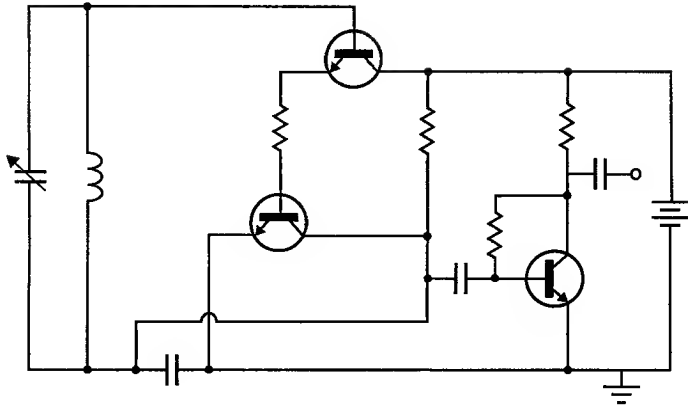
FIGURE 13.36: A circuit diagram for an AM radio.

A circuit intended for a printed circuit board can be naturally modeled by a graph. In Figure 13.37, we show a graph G that models the AM radio circuit in Figure 13.36. The vertices in G correspond to the circuit's components and junctions (locations where conducting pathways meet). The edges in G represent the conducting pathways in the circuit.
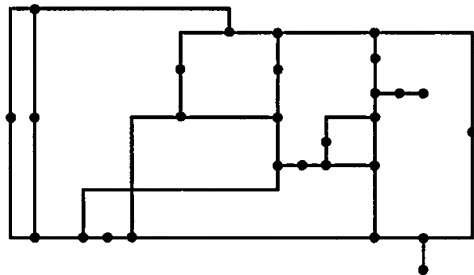
FIGURE 13.37: A graph model of the AM radio circuit.

In Figure 13.38, we illustrate an embedding of $K_{3,3}$ into the graph G. Therefore G is not embeddable in the plane, implying that every good drawing
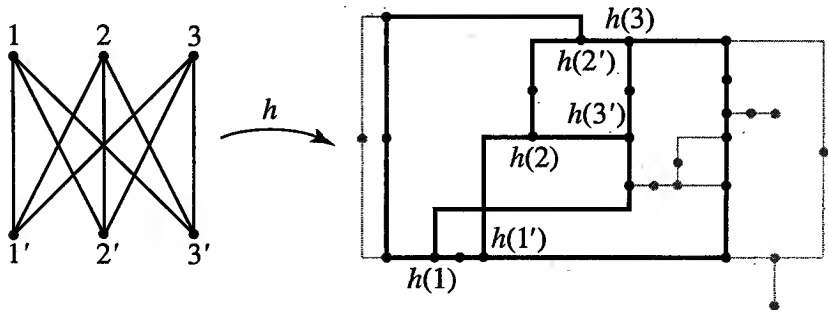
FIGURE 13.38: The graph $K_{3,3}$ embeds in G.

of G has at least one crossing point. It follows that there must be a conducting pathway overpass in any diagram of the circuit. Thus, construction of the AM radio circuit on a printed circuit board must involve drilling at least one pair of holes to allow a pathway to travel from, say, the top of the board to the bottom, under another pathway, and then back through to the top side of the board.

Mounting the components and etching the conducting pathways are the easy, low-cost aspects of manufacturing a printed circuit board. Production costs rise when a circuit requires drilling holes to use both sides of a board or layering a number of boards together to use multiple flat surfaces.

As we see in the AM radio circuit example, properties of graphs are helpful in planning a circuit layout and in determining how it can be constructed most economically on a printed circuit board. Crossing number and graph thickness are important graph properties that naturally play a role in printed circuit board design. The crossing number indicates the minimum number of crossings in a good drawing of a particular graph and therefore determines the minimum number of holes needed in a printed circuit board to pass circuit pathways under each other, ensuring no unwanted crossings. An alternative to drilling holes for crossings is to build the circuit in multiple layers, on the surfaces of a number of boards that are stacked together, insulated from each other, but electronically connected. For this approach, graph thickness provides an indication of the minimum number of surfaces needed to build a layered circuit. Thus, both crossing number and thickness can be valuable tools in designing and analyzing an electronic circuit so that it can be manufactured as economically as possible.

Exercises for Section 13.4

13.26. (a) In $\mathbb{R}$, assume that $a_1 < b_1 < c_1$ and $a_2 < b_2 < c_2$. Prove that there exists a homeomorphism $h : [a_1, c_1] \to [a_2, c_2]$ such that $h(b_1) = b_2$.

    (b) Discuss the role that part (a) plays in the proof of Theorem 13.26, where it is asserted that the homeomorphism h^* can be adjusted slightly to a homeomorphism h mapping all degree-2 vertices of G' to points of G that do not correspond to crossing points.

13.27. Prove that for a graph G the thickness and crossing number are related as follows: $\theta(G) \le 1 + \lceil \nu(G)/8 \rceil$ where $\lceil x \rceil$ denotes the least integer greater than or equal to x.

13.28. Prove that the bipartite graphs $K_{3,n}$, with $n \ge 3$, and $K_{4,m}$, with $m \ge 4$, have thickness 2.

13.29. **Prove Corollary 13.31:** For K_n and $K_{n,m}$ thickness is bounded from below as follows:

    (a) If $n \ge 3$, then $\theta(K_n) \ge \frac{n(n-1)}{6(n-2)}$.

    (b) If $n, m \in \mathbb{Z}_+$ and $n + m \ge 3$, then $\theta(K_{n,m}) \ge \frac{nm}{2(n+m-2)}$.

13.30. Prove that the graphs G_1 and G_2, constructed in the proof of Theorem 13.29, are planar.

13.31. Find a specific positive integer value k for which you can prove, using Corollary 13.31, that $\theta(K_{kn}) \ge n$ for all $n \ge 3$. (This demonstrates that there

are graphs with arbitrarily large thickness. Specifically, given any $n \geq 3$, the complete graph K_{kn} is a graph with thickness at least n.)

13.32. Consider the embedding of $K_{3,3}$ in the graph G illustrated in Figure 13.38. Sketch a diagram indicating how each individual edge in $K_{3,3}$ maps into G under the embedding.

13.33. Consider the circuit diagram in Figure 13.39. Create a graph model of the circuit and prove that the circuit cannot be manufactured on one side of a circuit board.
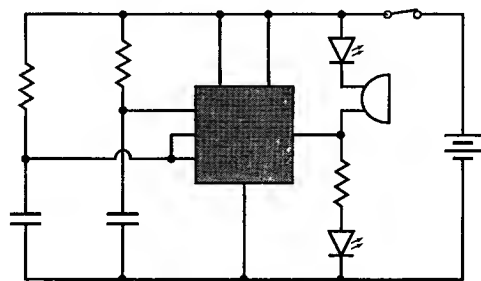
FIGURE 13.39: Prove that this circuit cannot be manufactured on one side of a circuit board.

13.34. Consider the circuit diagram in Figure 13.40. Prove that the circuit can be rearranged so that two, but not all three, of the crossings can be removed.
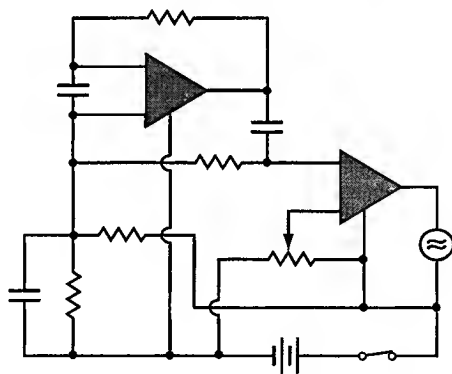
FIGURE 13.40: Prove that two, but not all three, of the crossings can be removed.

Chapter 14

Manifolds and Cosmology

One of the great intellectual advances achieved by humanity was the realization of the shape of the planet on which we reside. It took hundreds of thousands of years of human development before we were able to conclude that we live on a sphere.

Pythagoras posited a spherical earth as early as 500 B.C. A variety of observations suggested that the earth was curved. For example, masts were the last part of a ship to disappear on the horizon. Three hundred years later, Erastothenes used the shadows cast at noon in two different cities to estimate the curvature of the earth and from it, the earth's radius.

The fact that locally the surface of the earth is two-dimensional implies that it is a 2-manifold. We will define manifolds rigorously in the first section of this chapter. Other 2-manifolds we have already encountered include the torus, the Klein bottle, the projective plane, and the plane.

A priori knowing the local two-dimensionality of the earth's surface did not preclude the possibility that some 2-manifold other than a sphere corresponded to the earth. A torus earth certainly would have been interesting. (See Figure 14.1).

FIGURE 14.1: A torus Earth.

Since our realization of the shape of the earth, we have determined many other aspects of our universe. But one fundamental question remains unanswered. What is the shape of the universe within which we reside? Here, we mean the spatial universe, which can be taken as a spatial slice of four-dimensional space-time. We can picture the spatial universe as having been created at the time of the big bang and expanding in size since then. We would like to know what topological shape it is that is expanding.

In order to make sense of this question mathematically, we need to examine the fundamental nature of the spatial universe. What are its basic defining properties?

Notice that no matter where we have been in the universe so far, if we choose a given spot and travel out from it a short distance (say three feet) in all directions, we enclose a space that, for all intents and purposes, resembles a ball in three-dimensional Euclidean space. (See Figure 14.2.)

FIGURE 14.2: The set of points within three feet of a point is a ball.

The space that we live in appears to be locally three-dimensional. We call such an object a 3-manifold. Just as there are a variety of 2-manifolds, any one of which satisfies the same properties locally as does the surface of the earth, there are a variety of 3-manifolds, any of which is a candidate for the shape of the universe.

In this chapter, we discuss manifolds of dimensions one, two, and three in the first three sections, and in the last two sections we discuss how an understanding of 3-manifolds helps in the determination of the structure of the universe. We begin with the official definition of an n-manifold.

14.1 Manifolds

We want to capture the idea of a space that locally resembles real n-space $\mathbb{R}^n$.

DEFINITION 14.1. *An n-manifold is a Hausdorff topological space X with a countable basis such that each $x \in X$ has a neighborhood that is homeomorphic to the open n-ball.*

The fact that the neighborhoods are only required to be homeomorphic to the open n-ball gives us substantial leeway. A homeomorphism between a neighborhood and an open ball in $\mathbb{R}^n$ need not preserve distance; it only needs to deform the neighborhood to look like an open ball. As in Figure 14.3, the
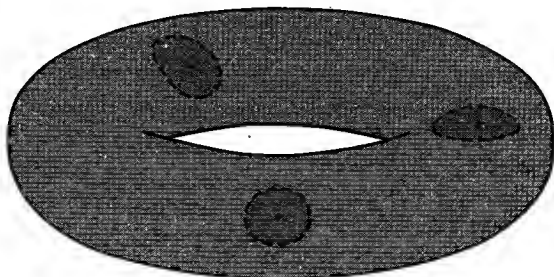
FIGURE 14.3 Each point on the torus has a neighborhood that is homeomorphic to an open disk.

torus has the property that each point has a neighborhood homeomorphic to an open disk in the plane, although the neighborhoods appear curved on the surface.

The two additional conditions that the space be Hausdorff and that it have a countable basis ensure that an n-manifold is a reasonably well-behaved space. (See Exercises 14.1 and 14.2.)

When considering n-manifolds for a given n, the most fundamental questions are the classification questions. Can we make a complete list of all n-manifolds? And, if such a list exists and we are given a particular n-manifold, is there an algorithm that will determine which manifold it is on the list?

This turns out to be a very difficult problem. It has been successfully completed for dimensions one and two. There has been substantial work in dimensions three and four, but a complete classification of 3-manifolds remains elusive. And it has been proven that there can never be an algorithm that determines whether or not two given 4-manifolds are homeomorphic. Thus the complete classification of 4-manifolds can never be attained. The same holds for the classification of n-manifolds for any $n \geq 4$. (See [Mar] and [Boo].)

Let us start with 1-manifolds.

DEFINITION 14.2. *A connected 1-manifold is called a* **curve.**

Each point in a curve has a neighborhood that is homeomorphic to the open interval $(-1, 1)$. There is a complete classification for 1-manifolds:

THEOREM 14.3. *Let X be a 1-manifold.*

    (i) If X is connected and compact, then X is homeomorphic to S^1.

    (ii) If X is connected and noncompact, then X is homeomorphic to $\mathbb{R}$.

    (iii) If X is disconnected, then X is the union of a countable collection of components, each of which is homeomorphic to either S^1 or $\mathbb{R}$.

Proofs of (i) and (ii) can be found in [Chr]. It is straightforward that (iii) follows from (i) and (ii). Specifically, if a 1-manifold X is not connected, then it is a union of components, each of which is a connected 1-manifold. Therefore, X is a union of mutually disjoint subspaces that are each homeomorphic to either S^1 or $\mathbb{R}$. Since a manifold must have a countable basis, there can be at most countably many such components in X.

Since we now have a good understanding of 1-manifolds, let us turn to 2-manifolds.

DEFINITION 14.4. *A connected 2-manifold is called a* **surface.**

The plane is a surface. It is connected and Hausdorff. Given a point $p \in \mathbb{R}^2$, the open ball of radius 1 centered at p is a neighborhood of p that is homeomorphic to the open disk. Furthermore, the plane has a countable basis given by the set of open balls of rational radius centered at points $x = (x_1, x_2)$ with both x_1 and x_2 rational.

Every connected open subset of the plane is also a surface. In this section and the next, our primary focus is on compact surfaces.

THEOREM 14.5. *The sphere, the torus, the Klein bottle, and the projective plane are compact surfaces.*

Proof. We discuss the proof for the Klein bottle. The proof is essentially the same for the other three spaces. We saw earlier that the Klein bottle is connected and compact, being the image of a connected and compact space under a quotient mapping.

Assume that we have a Klein bottle that is constructed from the square $[0, 1] \times [0, 1]$ by gluing the edges in the usual way. As illustrated in Figure 14.4, a point x in the interior of the square has an open-disk neighborhood obtained by taking an open ball of appropriately small radius. For a point y on an edge of the square (but not a vertex) and the point y' on the opposite edge with which it is identified, a half-disk neighborhood of y and a half-disk neighborhood of y' glue together to form an open-disk neighborhood of the resultant point in the Klein bottle. Finally, a vertex of the square has a quarter-disk neighborhood. The vertex is identified with the other three vertices, each of which has a similar quarter-disk neighborhood. These four quarter-disk neighborhoods glue together to form an open-disk neighborhood of the resultant point in the Klein bottle. Hence, every point in the Klein bottle has an open-disk neighborhood.
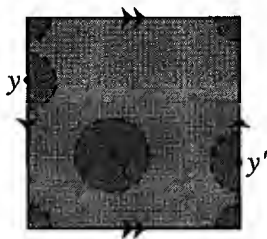
FIGURE 14.4: Each point in the Klein bottle has a neighborhood homeomorphic to an open disk.

The Klein bottle is Hausdorff. Given two points in the Klein bottle, we can separate them using sufficiently small open-disk neighborhoods, glued half-disk neighborhoods, or glued quarter-disk neighborhoods, as needed.

Finally, the square, $[0, 1] \times [0, 1]$, has a countable basis $\mathcal{B}$ obtained by taking the intersection of the square with every open ball in the plane having rational radius and centered at points with rational coordinates. From $\mathcal{B}$, take the balls that lie in the interior of the square, the half balls that are centered on the edges and do not contain vertices, and the quarter balls centered on the vertices, then map them into the Klein bottle, gluing the half balls and quarter balls appropriately. The result is a countable basis for the Klein bottle. ∎

We would like to use various attributes of surfaces to distinguish between them. One of these is the concept of orientability. To get a handle on orientability, we start with the prototype nonorientable space, the Möbius band.

A Möbius band has the property that it can reverse orientation in the following sense: Choose a start point on the core curve of the Möbius band M, and take a clockwise rotation in a neighborhood of that point, as illustrated on M in Figure 14.5. Now, let us travel around the core curve, pulling the clockwise rotation along with us. By the time we have returned to our starting point, our clockwise rotation has been reversed into a counterclockwise rotation. We have reversed our orientation.

On the other hand, consider the clockwise rotation on the annulus A in Figure 14.5. If we pull the rotation around the core curve, it does not get reversed.
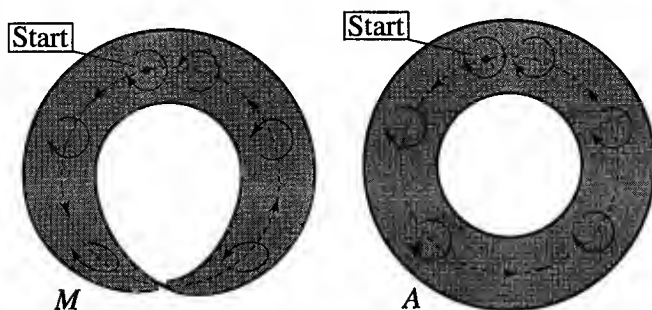
FIGURE 14.5: Pulling directions of rotation around the Möbius band and the annulus.

The property of possessing a curve that reverses orientation is held by every surface that contains an embedded copy of the Möbius band. This motivates the following definition:

DEFINITION 14.6. *A surface that contains an embedded Möbius band is called **nonorientable**. A surface that does not contain an embedded Möbius band is called **orientable**.*

As indicated by the darker shaded regions in the Klein bottle and in the projective plane in Figure 14.6, both of these surfaces contain an embedded Möbius band, and therefore they are both nonorientable surfaces.
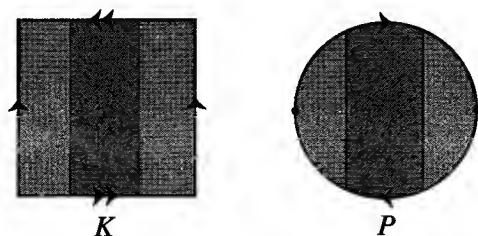
FIGURE 14.6: The Klein bottle and the projective plane contain embedded Möbius bands.

The next theorem indicates that orientability is a topological invariant. Therefore we can use orientability to help distinguish between surfaces.

THEOREM 14.7. *Let S_1 and S_2 be homeomorphic surfaces. Then S_1 is orientable if and only if S_2 is orientable.*

   Proof. See Exercise 14.4.   ■

   In general, it is difficult to prove that a surface is orientable by directly proving that it does not contain an embedded Möbius band. Therefore we take a different approach to demonstrate orientability, one that involves triangulations—decompositions of a surface into triangles. After we discuss triangulations and some helpful related results, we will come back to the matter of orientability.

   Let τ be a triangular region in the plane, and let S be a surface. If $f : \tau \to S$ is an embedding, then we say that $f(\tau)$ is a **triangle in** S. Also, the images under f of the edges and vertices of τ are called the **edges** and **vertices**, respectively, of $f(\tau)$. (See Figure 14.7.)
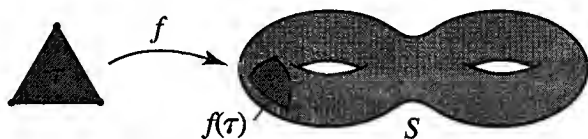
FIGURE 14.7: A triangle in the surface S.

DEFINITION 14.8. *A **triangulation** T of a compact surface S is a collection of finitely many triangles in S that covers S and is such that any two triangles in T either do not intersect, intersect in a vertex of both triangles, or intersect in an edge of both triangles.*

   In Figure 14.8, we show triangulations of the sphere, torus, Klein bottle, and projective plane.
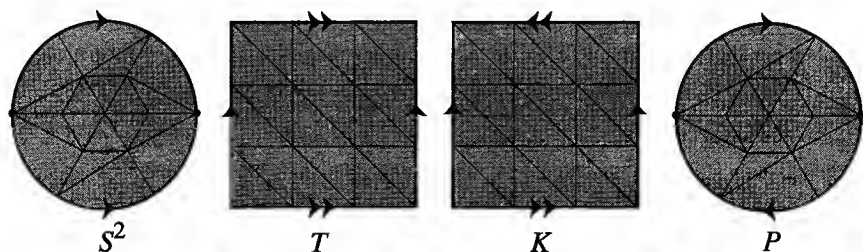
FIGURE 14.8: Triangulations of compact surfaces.

THEOREM 14.9. *Every compact surface has a triangulation.*

This result was first proved by Tibor Radó in 1925. The proof is quite technical, and therefore we do not pursue it here. Proofs can be found in Radó's original paper, [Rad], and, more recently, in the text [Moi].

The next definition and theorem will be important for our work with triangulations in this and in the next section.

DEFINITION 14.10. *Let T be a triangulation of a compact surface S. Then*

(i) *A triangulation T' is called a* **subdivision** *of T if every triangle in T can be expressed as a union of triangles in T'.*

(ii) *A triangulation T^* is said to be* **equivalent** *to T if there is a homeomorphism $h : S \to S$ that maps the triangles in T homeomorphically to the triangles in T^*.*

THEOREM 14.11. *Let T_1 and T_2 be triangulations of a compact surface S. There exist triangulations T_1' and T_2' that are equivalent to each other and are subdivisions of T_1 and T_2, respectively.*

Proof. Suppose we have triangulations T_1 and T_2 of a compact surface S. The triangulations are said to be in general position if the following conditions hold:

(i) Every vertex of each triangulation lies in the interior of a triangle of the other triangulation;

(ii) Every edge of each triangulation intersects every edge of the other triangulation in at most finitely many points.

If T_1 and T_2 are in general position, then we can construct a triangulation T' that is a subdivision of both T_1 and T_2. The following steps are used to construct T'; we illustrate them in Figure 14.9:

(i) The vertices in T_1 and the vertices in T_2 become vertices in T'.

(ii) Each point of intersection between an edge in T_1 and an edge in T_2 becomes a vertex in T'.

(iii) Each edge that is in T_1 or in T_2 either becomes an edge in T' or is subdivided into edges in T' by vertices from step (b).

(iv) New vertices are added in edges to form triangles out of regions that are bounded by only two edges.

(v) New edges are added in regions bounded by more than three edges to subdivide such regions into triangles.
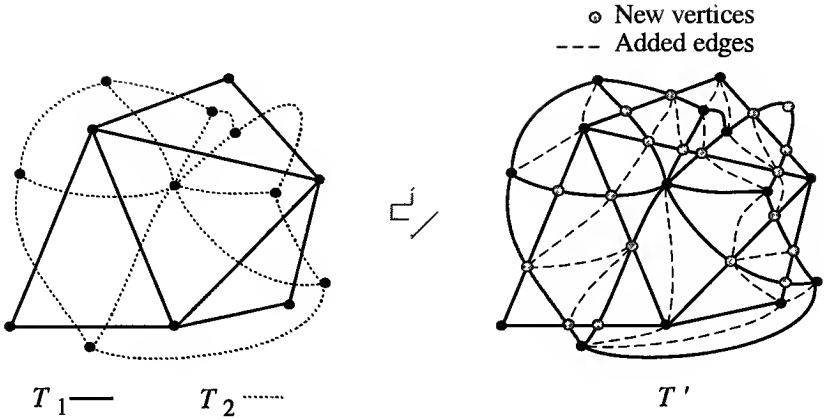
FIGURE 14.9: Constructing a subdivision T' from triangulations T_1 and T_2.

The resulting triangulation T' is a subdivision of both T_1 and T_2, and therefore the theorem holds in the case that T_1 and T_2 are in general position.

Next, suppose that T_1 and T_2 are not in general position. Then we can make small adjustments to one of the triangulations, say T_1, to obtain a new triangulation T_1^* that is equivalent to T_1 but is such that T_1^* and T_2 are in general position. That we can do this is plausible, but it is a technical result that takes a lot of detail to develop formally. Thus let T_1^* be such a triangulation, and let $f : S \rightarrow S$ be a homeomorphism mapping T_1 to T_1^*. Since T_1^* and T_2 are in general position, we can, as before, find a subdivision T_2' of both T_1^* and T_2. Then we can map T_2' by f^{-1} to a subdivision T_1' of T_1. In this way, we have obtained subdivisions T_1' and T_2' of T_1 and T_2, respectively, such that T_1' and T_2' are equivalent. ∎

As illustrated in Figure 14.10, a triangle can be assigned a direction of rotation that can serve as the direction of rotation at each of its points and that does not result in the orientation-reversal issues encountered on the Möbius band.
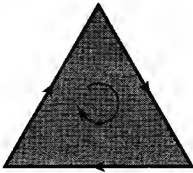
FIGURE 14.10: A direction of rotation on a triangle.

Also, as shown in Figure 14.10, a direction of rotation on a triangle induces a direction on each of its edges. This leads us to the following definition:

DEFINITION 14.12. *Let S be a compact surface and T be a triangulation of S. The triangulation T can be* **coherently oriented** *if each triangle in T can be assigned a direction of rotation such that if two triangles intersect in an edge, then the directions the edge inherits from each of the triangles are opposite each other. (See Figure 14.11.)*
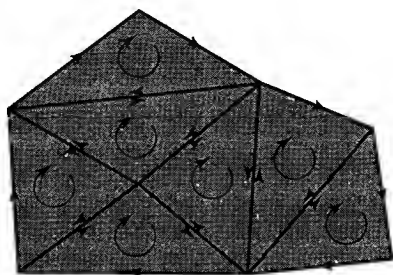
FIGURE 14.11: Triangles in a coherently oriented triangulation.

In this definition, the condition that the inherited edge directions do not match ensures that we can slide an orientation from one triangle across an edge to an adjacent triangle and have the corresponding orientations agree.

In Figure 14.12 we depict coherently oriented triangulations of the sphere and the torus. In Figure 14.8 the triangulation of the projective plane appears similar to the triangulation of the sphere. However, the coherent orientation of the triangulation of the sphere does not carry over to the projective plane because when we glue the top and bottom edges of the disk to obtain the projective plane, there are triangles whose edges are glued together so that the orientations do not properly match. A similar situation occurs with the triangulations of the torus and the Klein bottle.
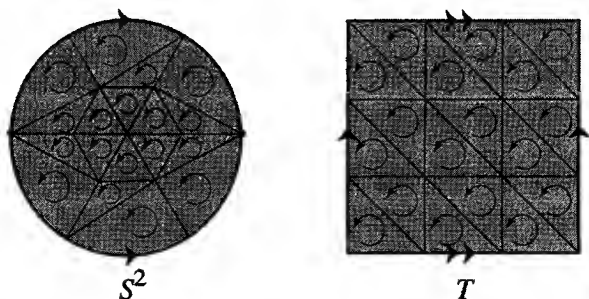
FIGURE 14.12: Coherently oriented triangulations of the sphere and torus.

The following theorem establishes the equivalence between being orientable and having a coherent orientation:

THEOREM 14.13. *A compact surface S is orientable if and only if it has a triangulation that can be coherently oriented.*

Proof. We first show that if one triangulation of a compact surface S can be coherently oriented, then every triangulation of S can be coherently oriented. Specifically, let T_1 and T_2 be triangulations of S, and suppose that T_1 is coherently oriented. By Theorem 14.11, there exist subdivisions T_1' and T_2' of T_1 and T_2, respectively, such that T_1' and T_2' are equivalent. It is evident that

(i) Subdivision T_1' inherits a coherent orientation from the coherent orientation of T_1.

(ii) Triangulation T_2' obtains a coherent orientation as a result of its equivalence to T_1'.

(iii) Triangulation T_2 obtains a coherent orientation from the coherent orientation of its subdivision T_2'.

Now, suppose that S is a nonorientable compact surface and therefore contains an embedded Möbius band, M. We need to show that every triangulation of S cannot be coherently oriented. By the first part of the proof, it suffices to show that there exists one triangulation of S that cannot be coherently oriented. Let T be a triangulation of S. We can adjust and subdivide T, in a manner similar to the construction in the proof of Theorem 14.11, so that in the resulting triangulation T' of S there is a collection of triangles, T_M', whose union is the embedded Möbius band M. The fact that we can reverse a direction of rotation by pulling it around the core curve in a Möbius band implies that the triangles in T_M' cannot be coherently oriented. Thus, the triangulation T' of S cannot be coherently oriented.

Next, suppose that S is orientable. We need to show that it has a triangulation that can be coherently oriented. Suppose it does not, and let T be a triangulation of S. Then T cannot be coherently oriented. Starting with any triangle in T, choose a direction of rotation on it. Extend this direction of rotation to the adjacent triangles in a coherent manner. Continue this process over the triangles in T. Because T cannot be coherently oriented, at some point we end up with a direction of rotation on a triangle that is not coherent with a rotation already assigned to an adjacent triangle. We can then find a sequence of distinct triangles $\tau_1, \ldots, \tau_n$ such that

(i) For each $i = 1, \ldots, n-1$, τ_i and τ_{i+1} have an edge in common and are coherently oriented with each other;

(ii) Triangles τ_n and τ_1 have an edge in common but the rotation on τ_n is not coherent with the rotation on τ_1.

Consider the strip running across each triangle τ_i as shown shaded in Figure 14.13. If we regard the strip as being constructed by gluing the triangles together, one at a time, then when we glue τ_n to τ_1, the strip is glued with a half twist, making it a Möbius band. Therefore there is a Möbius band embedded in S, contradicting the assumption that S is

orientable. It follows that S has a triangulation that can be coherently oriented.   ■
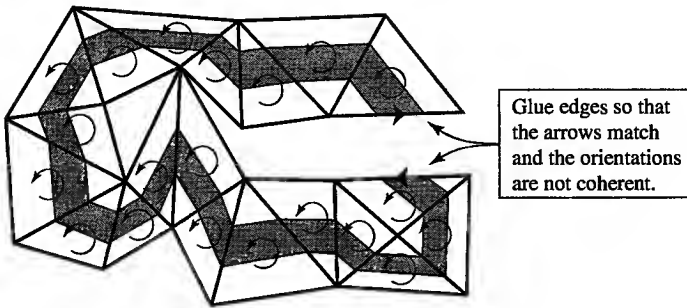
FIGURE 14.13: The shaded strip glues to form a Möbius band in the surface.

We pointed out previously that the sphere and the torus each have triangulations that can be coherently oriented. (See Figure 14.12.) Therefore Theorem 14.13 implies that the sphere and torus are orientable. Since orientability is a topological property, it follows that we can distinguish these two surfaces from the Klein bottle and the projective plane. Specifically, neither the sphere nor the torus is homeomorphic to either the Klein bottle or the projective plane because the former two are orientable while the latter two are not. As in this example, orientability is a fundamental property that is used to distinguish between compact surfaces. In the next section, we use it, along with the Euler characteristic, to distinguish between all compact surfaces.

Up to this point, we have introduced a somewhat limited collection of compact surfaces, but we can dramatically expand that collection using the connected-sum operation, defined as follows:

DEFINITION 14.14. *Given two surfaces S_1 and S_2, the **connected sum of** S_1 **and** S_2, denoted $S_1\#S_2$, is the surface obtained by removing the interior of a disk from each surface and gluing the two circle boundaries together via a quotient map. (See Figure 14.14.)*
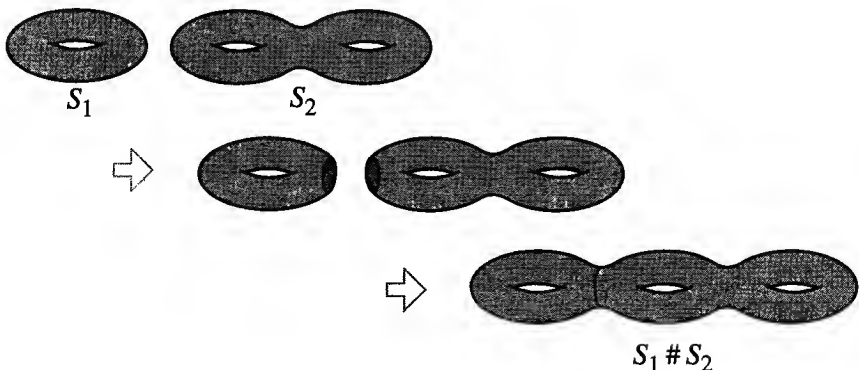
FIGURE 14.14: The connected sum of two surfaces.

The result of the connected-sum operation is a surface. (See Exercise 14.9.) Although we do not prove it here, up to homeomorphism the connected-sum operation is independent of the particular disks whose interiors we remove from the surfaces and the particular choice of a gluing we use on the two boundary circles.

Notice that when we take the connected sum of the sphere S^2 with any other surface S, the result is homeomorphic to S. (See Figure 14.15.) If we think of connected sum as an additive operation on surfaces, S^2 acts like the identity.

$$S^2 \# S \qquad \cong \qquad S$$
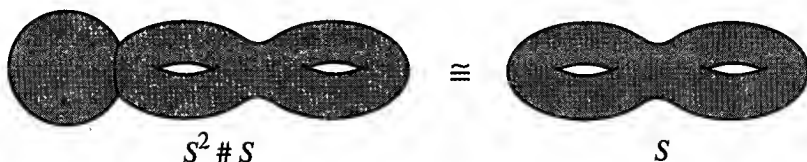
FIGURE 14.15: $S^2\#S$ is homeomorphic to S.

DEFINITION 14.15. *The connected sum of a n tori, denoted* nT, *is called a* **genus n surface**. *A sphere is called a* **genus 0 surface**. *(See Figure 14.16.)*

$$n = 0 \qquad n = 1 \qquad n = 2$$

$$n = 3$$
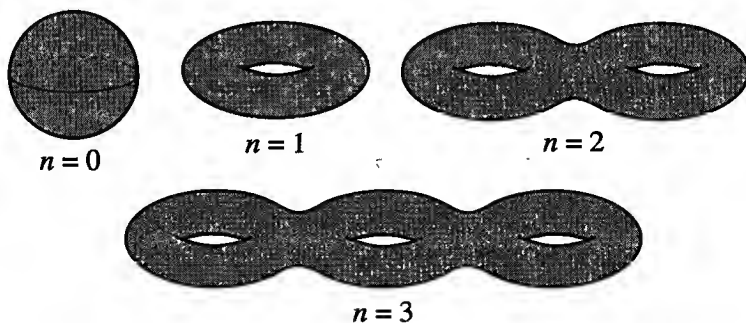
FIGURE 14.16: The genus n surfaces, with $n = 0, 1, 2, 3$.

The genus n surface is a compact surface for all integers $n \geq 0$. Furthermore, by the next theorem it follows that the genus n surface is orientable.

THEOREM 14.16. *Let* S_1 *and* S_2 *be compact surfaces. Then* $S_1\#S_2$ *is orientable if and only if both* S_1 *and* S_2 *are orientable.*

Proof. See Exercise 14.10. ∎

Exercises for Section 14.1

14.1. The extra-point line was introduced and shown to be non-Hausdorff in Exercise 7.18. The underlying set is $X = \mathbb{R} \cup \{p_e\}$, where p_e is an extra point, not contained in $\mathbb{R}$. The topology on X is defined via the basis $\mathcal{B}$ consisting of all intervals $(a, b) \subset \mathbb{R}$ and all sets of the form $(c, 0) \cup \{p_e\} \cup (0, d)$ for $c < 0$ and $d > 0$.
   (a) Show that every point in X has a neighborhood homeomorphic to an open interval.
   (b) Show that X has a countable basis.
   (The extra-point satisfies two of the conditions needed to be a 1-manifold, but it is not Hausdorff and therefore is not a 1-manifold.)

14.2. Show that if $\mathbb{R}_d$ is $\mathbb{R}$ with the discrete topology, then $\mathbb{R}_d \times \mathbb{R}$ satisfies all of the properties of being a 1-manifold except that it does not have a countable basis.

14.3. Let M be an n-manifold and assume $x \in M$. Show that for each neighborhood U of x there is an open set B such that $x \in B \subset U$ and B is homeomorphic to the open n-ball.

14.4. **Prove Theorem 14.7:** Let S_1 and S_2 be homeomorphic surfaces. Then S_1 is orientable if and only if S_2 is orientable.

14.5. Although the Möbius band is not a surface, we can triangulate it as shown in Figure 14.17. Choose a direction of rotation on one of the triangles, and prove that it cannot be extended to a coherent orientation for the whole triangulation.
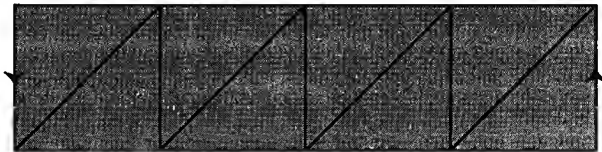
FIGURE 14.17: A triangulation of the Möbius band.

14.6. Show that every manifold is regular and therefore metrizable by the Urysohn Metrization Theorem (Theorem 5.24). (Hint: Use Exercise 14.3.)

14.7. Create an example of a topological space that satisfies all of the conditions for being a 2-manifold except that it is not Hausdorff. (Hint: Consider Exercise 14.1.)

14.8. Let M be a noncompact n-manifold.
   (a) Show that M is locally compact. (Therefore, the one-point compactification of M yields a compact Hausdorff space Y that contains M as a subspace and is such that $\text{Cl}(M) = Y$.)
   (b) Give an example of a 2-manifold M such that the one-point compactification of M is not a 2-manifold.

14.9. Show that the connected sum of two compact surfaces is a compact surface.

14.10. **Prove Theorem 14.16:** Let S_1 and S_2 be compact surfaces. Then $S_1 \# S_2$ is orientable if and only if both S_1 and S_2 are orientable.

14.2   *Euler Characteristic and the Classification of Compact Surfaces*

In this section, we continue our investigation of compact surfaces. We introduce the Euler characteristic, prove that it is an invariant for compact surfaces, and then use it—in conjunction with orientibility—to distinguish between compact surfaces.

Given a triangulation of a compact surface S, we can construct a polygon with pairs of edges to be identified that yields the surface. We begin with one of the triangles and perform a single-edge identification with another triangle. The result is topologically a disk. If there are remaining triangles, then because S is a surface, at least one will glue to this disk along an edge on its boundary. Again, the result is topologically a disk. We continue this process until all triangles have been attached, resulting in a disk with pairs of edges in its boundary still to be identified. This is called a **polygonal representation** of S. The representations of the torus and the Klein bottle as squares with their edges glued together are polygonal representations. We saw other polygonal repɾɔsentations of surfaces in Section 3.4.

EXAMPLE 14.1. In Figure 14.18 we see a polygonal representation of $T\#K$, the connected sum of a torus and a Klein bottle. In the illustration of the resulting surface, we show how the glued edges appear.
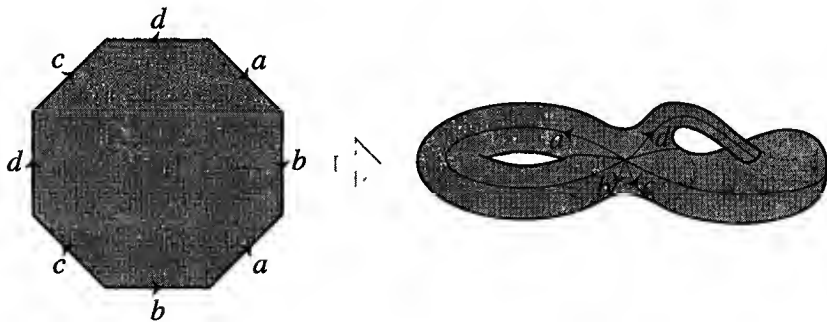
FIGURE 14.18: A polygonal representation of $T\#K$.

Given a triangulation of a compact surface, we define its Euler characteristic as follows:

DEFINITION 14.17. *Let T be a triangulation of a compact surface S. We define the **Euler characteristic of** T by $\chi(T) = V - E + F$ where V is the number of vertices, E is the number of edges, and F is the number of triangles in the triangulation.*

We use F for the number of triangles because we also think of them as faces, as if the triangulated compact surface is made up of triangular faces. The

expression $V - E + F$ appears in both the Euler characteristic formula and Euler's formula for planar graphs (Theorem 13.18). We discuss the connection between the two formulas later in this section.

It turns out that the Euler characteristic is the same for every triangulation of a given compact surface. We express this in the following important theorem, which will be very useful in distinguishing compact surfaces:

THEOREM 14.18. *Let T_1 and T_2 be triangulations of a compact surface S. Then $\chi(T_1) = \chi(T_2)$.*

> **Proof.** Let S be a compact surface, and let T_1 and T_2 be triangulations of S. First, if T_1 and T_2 are equivalent, then it is straightforward that their Euler characteristics are the same.
>
> On the other hand, if T_1 and T_2 are not equivalent, then by Theorem 14.11 there are equivalent triangulations T_1' and T_2' that are subdivisions of T_1 and T_2, respectively. We claim that the Euler characteristic is preserved under subdivision. Given the claim, we then have $\chi(T_1) = \chi(T_1'), \chi(T_2) = \chi(T_2')$, and, since T_1' and T_2' are equivalent, $\chi(T_1') = \chi(T_2')$. It follows that $\chi(T_1) = \chi(T_2)$ in this case as well.
>
> Thus we need to show that the Euler characteristic of a triangulation T is preserved under subdivision. To do so, we think of T as a topological graph G in S, made up of the vertices and edges of T. We can construct a subdivision T' of T by applying a sequence of operations to G, generating a sequence of topological graphs $G = G_0, \ldots, G_n$, where G_n is the topological graph corresponding to the subdivision T'.
>
> The operations are as follows; we illustrate them in Figure 14.19:
>
> > (i) Add a new vertex that lies on an edge in G_i to obtain G_{i+1}. As a result, the edge in G_i is subdivided into two new edges in G_{i+1}.
> >
> > (ii) Add a new vertex and a new edge to G_i to obtain G_{i+1}. The added vertex and edge lie in the complement of G_i, and the new edge connects the new vertex to a vertex in G_i.
> >
> > (iii) Add a new edge to G_i to obtain G_{i+1}. The added edge lies in the complement of G_i, and it connects two vertices in G_i.
>
> To show that the Euler characteristic is preserved under subdivision, we show that a sequence of these operations does not change the Euler characteristic. First, when a topological graph G is associated to a triangulation T of S, then the vertices and edges of G correspond to the vertices and edges of T, and the components of the complement of G in S correspond to the triangles in T. Therefore, if we let V represent the number of vertices in G, E the number of edges, and F the number of components of the complement of G in S, then $V - E + F$ is the Euler characteristic. After each of the operations is applied to a topological graph, we do not necessarily have a topological graph that is associated with a triangulation, but we can still compute $V - E + F$ for it. We claim
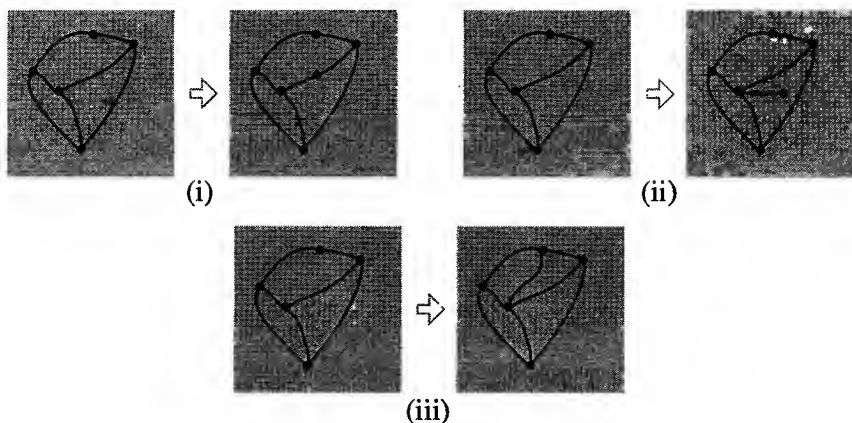
FIGURE 14.19: The operations on topological graphs for subdividing triangulations.

that this value does not change under each of the operations, and therefore it follows that the Euler characteristic is unchanged under subdivision. Let us look at the effect of the operations on $V - E + F$:

Operation (i):   This operation adds a vertex to G_i and subdivides an edge into two new edges. Therefore we have a net gain of one vertex and one edge, resulting in no change to $V - E + F$.

Operation (ii):   This operation adds a vertex and an edge to G_i and leaves the number of components of the complement unchanged. Therefore we have a net gain of one vertex and one edge, resulting in no change to $V - E + F$.

Operation (iii):   This operation adds an edge to G_i and splits one component of the complement into two new components. Therefore we have a net gain of one edge and one component of the complement, resulting in no change to $V - E + F$.

It follows that the Euler characteristic is preserved under subdivision, and therefore we obtain the desired result: If T_1 and T_2 are triangulations of a compact surface S, then $\chi(T_1) = \chi(T_2)$.   ∎

Although intuitively apparent, it is nontrivial that operation (ii) in the foregoing proof leaves unchanged the number of components of the complement of the topological graph while operation (iii) splits one component into two new ones. These results follow from Theorem 13.17, a result that we did not prove but that generalizes the Jordan Curve Theorem (Theorem 11.2) and Theorem 11.10.

Operations (i)–(iii) from the proof of Theorem 14.18 enable us to carry out the five steps that were used in the proof of Theorem 14.11 to construct a common subdivision of two triangulations in general position. Let T_1 and T_2 be two such triangulations. If we take T_1, then we can use operations (i)–(iii) to add in all of the vertices and edges (or subdivided edges) from T_2 to

superimpose the two triangulations. This essentially covers construction steps (i)–(iii) in the proof of Theorem 14.11. Then we can use operations (i)–(iii) again to insert vertices and edges (as described in steps (iv) and (v) in the proof of Theorem 14.11) to make triangles out of regions that are not already triangles.

Now that we have Theorem 14.18, we can define the **Euler characteristic of a compact surface** S to be the Euler characteristic of any triangulation of S. We denote the Euler characteristic of S by $\chi(S)$.

It is clear that if T is a triangulation of S, and $f : S \to S'$ is a homeomorphism, then f naturally maps T to a triangulation T' of S'. Furthermore, the Euler characteristic of T' equals that of T. Thus we have the following theorem, which establishes that the Euler characteristic is an invariant for compact surfaces:

THEOREM 14.19. *If S and S' are homeomorphic compact surfaces, then* $\chi(S) = \chi(S')$.

EXAMPLE 14.2. From the triangulations in Figure 14.8, we see that $\chi(S^2) = 2$, $\chi(T) = 0$, $\chi(K) = 0$, and $\chi(P) = 1$.

These Euler characteristic values, along with orientability, enable us to distinguish these four spaces. By Theorem 14.19, it follows that S^2 is not homeomorphic to T, K, or P, that T is not homeomorphic to P, and that K is not homeomorphic to P. The only distinction not made here by the Euler characteristic is between T and K, but we already made that distinction in the last section, using orientability.

The topological invariance of the Euler characteristic (Theorem 14.18) and Euler's formula for planar graphs (Theorem 13.18) are closely related results. Theorem 14.18 applies to compact surfaces, but only addresses triangulations, while Theorem 13.18 only applies to the plane and the sphere, but addresses general topological graphs. Both results imply that the Euler characteristic of the sphere is 2.

As we have done with S^2, T, K, and P, we can use the Euler characteristic and orientability to distinguish all compact surfaces. The following theorem is an excellent example of the kind of result we seek. It gives a complete classification of compact surfaces.

THEOREM 14.20. *Every compact surface is homeomorphic to exactly one of S^2, $T \# T \# \ldots \# T$, or $P \# P \# \ldots \# P$.*

The compact surfaces listed in Theorem 14.20 are illustrated in Figure 14.20. Just as nT represents the connected sum of n tori, we use nP to denote the connected sum of n projective planes.

In order to prove Theorem 14.20, it is necessary to show that every compact surface is homeomorphic to at least one of S^2, nT, or nP, and that no two of the surfaces among S^2, nT, and nP are homeomorphic. Later in this section we use the Euler characteristic and orientability to show that no two of the listed surfaces are homeomorphic.
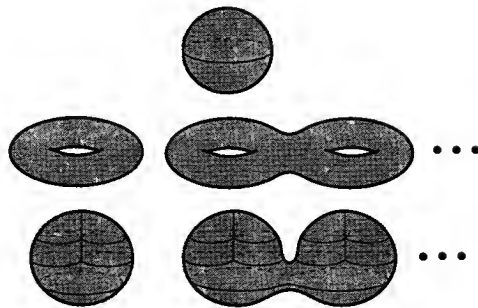
FIGURE 14.20: The complete collection of compact surfaces.

We do not prove that every compact surface is homeomorphic to at least one of the listed surfaces. Proofs can be found in [Mun] and [Mas]. The idea is that given a compact surface, we can perform a series of cut-and-paste operations on it to show that it is in the form of one of the listed surfaces. We introduced cutting and pasting in Section 3.4. We present another cutting-and-pasting example in Example 14.3.

Considering all of the compact surfaces that we can construct using connected sums, it is surprising that they all fall into one of the types listed in Theorem 14.20. For example, the Klein bottle K is a compact surface that does not appear to be among the listed surfaces. However, we showed in Example 3.25 that $K \cong P\#P$. Of course, $P\#P$ is one of the listed surfaces, and therefore so is the Klein bottle.

EXAMPLE 14.3. Another compact surface that does not appear to be among the surfaces listed in Theorem 14.20 is $T\#P$. This is a nonorientable surface. Therefore, by Theorem 14.20, it must be a connected sum of projective planes. In fact, $T\#P \cong P\#P\#P$. In Figure 14.21 we show by cutting and pasting that $T\#P \cong K\#P$. Since $K \cong P\#P$, it follows that $T\#P \cong K\#P \cong P\#P\#P$.

The relationship $T\#P \cong K\#P$ from Example 14.3 demonstrates that there is no cancellation law for the connected sum of surfaces. Such a cancellation law would imply that T is homeomorphic to K given that $T\#P$ is homeomorphic to $K\#P$. However, T is not homeomorphic to K since T is orientable but K is not.

The relationships $P\#P \cong K$ and $T\#P \cong K\#P$ yield the following characterization of connected sums of projective planes:

THEOREM 14.21. *For every integer $n \geq 3$,*

$$nP \cong \begin{cases} \frac{n-1}{2}T\#P & \text{if } n \text{ is odd}, \\ \frac{n-2}{2}T\#K & \text{if } n \text{ is even}. \end{cases}$$

Proof. See Exercise 14.11. ∎
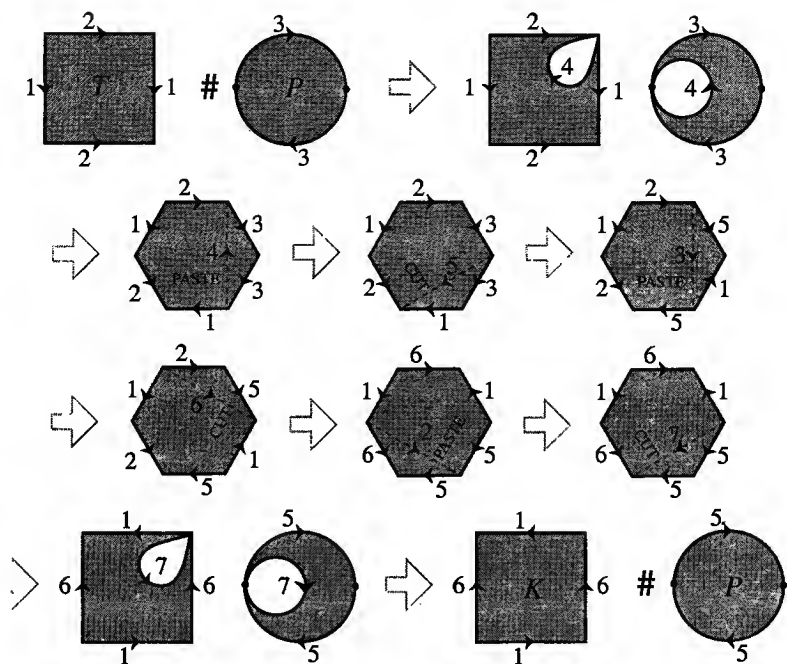
FIGURE 14.21: Using cutting and pasting to show that $T\#P$ is homeomorphic to $K\#P$.

The following alternative form of the classification theorem for compact surfaces is a straightforward consequence of Theorems 14.20 and 14.21:

THEOREM 14.22. *Every compact surface is homeomorphic to exactly one of S^2, nT, $nT\#K$, or $nT\#P$.*

We allow the possibility of no Ts in the latter two connected sums in Theorem 14.22 to include K and P in our list. In Figure 14.22, we illustrate the compact surfaces listed in Theorem 14.22.
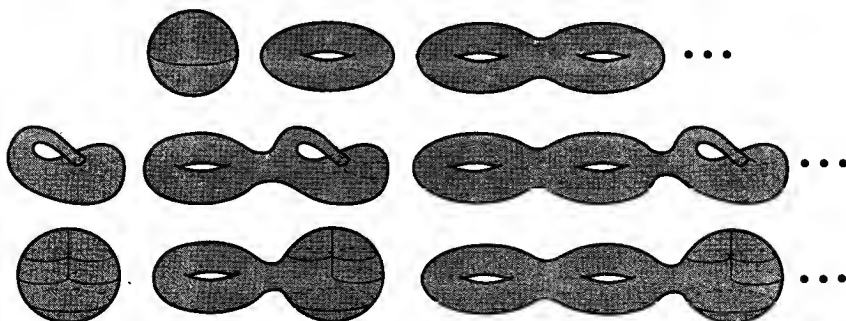
FIGURE 14.22: The complete collection of compact surfaces.

As demonstrated with the sphere, torus, Klein bottle, and projective plane in Example 14.2, we can use orientibility and the Euler characteristic to distinguish between different compact surfaces. In order to distinguish between all of the compact surfaces, we need to understand the effect of the connected sum operation on the Euler characteristic.

LEMMA 14.23. *If S_1 and S_2 are compact surfaces, then*

$$\chi(S_1 \# S_2) = \chi(S_1) + \chi(S_2) - 2.$$

Proof. Since the Euler characteristic is independent of the triangulation, we choose triangulations of S_1 and S_2 so that the disks whose interiors we remove in the process of taking the connected sum are triangles in the triangulations. Then the connected-sum operation corresponds to removing the interior of a triangle from each surface and identifying the boundaries of the triangles with each other, gluing vertices to vertices and edges to edges. But each boundary consists of three vertices and three edges. Hence, in the process of combining the triangulations of S_1 and S_2 to obtain a triangulation of $S_1 \# S_2$, we lose three vertices, three edges, and two faces. The resulting Euler characteristic is then $\chi(S_1) + \chi(S_2) - 3 + 3 - 2$. Therefore, $\chi(S_1 \# S_2) = \chi(S_1) + \chi(S_2) - 2$. ∎

THEOREM 14.24. *For the compact surfaces nT and nP, the Euler characteristic is given by $\chi(nT) = 2 - 2n$ and $\chi(nP) = 2 - n$.*

Proof. See Exercise 14.12. ∎

By our results on orientability from the previous section, including Theorem 14.16, it follows that S^2 and nT are orientable while nP is not. Furthermore, since $\chi(S^2) = 2$ and $\chi(nT) = 2 - 2n$, it follows that no two of the compact surfaces among S^2 and nT have the same Euler characteristic. Similarly, no two of the compact surfaces nP have the same Euler characteristic. Thus, given any two of the compact surfaces listed in Theorem 14.20, either they do not have the same Euler characteristic or one is orientable and the other is not. Either way, we can conclude that no two of the compact surfaces listed in Theorem 14.20 are homeomorphic. We can draw the same conclusion for the compact surfaces listed in Theorem 14.22. This yields the following powerful result:

THEOREM 14.25. *Two compact surfaces are homeomorphic if and only if they have the same Euler characteristic and are either both orientable or both nonorientable.*

Theorems 14.20 and 14.25 finish off the classification problem for compact surfaces. We have a complete list of possibilities, with no repetition, and a simple algorithm (compute the Euler characteristic and identify whether or not the surface is orientable) for determining which compact surface listed in Theorem 14.20 corresponds to any given surface.

EXAMPLE 14.4. To see how the algorithm can be used to identify a compact surface, consider the surface S given by the connected sum $K\#K\#P\#P\#T\#T$. It follows that S is nonorientable since there is a nonorientable surface in the connected sum out of which S is composed. Using Lemma 14.23 and the Euler characteristics for K, T, and P, we find that $\chi(S) = -8$. Among the compact surfaces listed in Theorem 14.20, the nonorientable surface with Euler characteristic -8 is $10P$, the connected sum of 10 projective planes, and therefore S is homeomorphic to $10P$.

We would also like to consider a disk or a Möbius band as a surface, however, the points on their edges do not have neighborhoods homeomorphic to the open 2-ball. But they do have neighborhoods homeomorphic to half the ball. Thus, let H^n be the subspace of the open n-ball $\mathring{B}^n$ given by $H^n = \{(x_1, x_2, \ldots, x_n) \in \mathring{B}^n \mid x_n \geq 0\}$. We define a broader class of manifolds as follows.

DEFINITION 14.26. *An n-manifold M with boundary is a Hausdorff space with a countable basis such that every point either has a neighborhood home-omorphic to the open n-ball $\mathring{B}^n$ or has a neighborhood homeomorphic to H^n, and the set of points of the latter type is nonempty.*

The set of points that have a neighborhood homeomorphic to $\mathring{B}^n$ is called the **interior** *of the manifold, and the set of points that have no neighborhood homeomorphic to $\mathring{B}^n$, but have a neighborhood homeomorphic to H^n, is called the* **boundary** *of the manifold. The boundary of M is denoted ∂M.*

A connected 2-manifold with boundary is called a **surface with boundary**.

EXAMPLE 14.5. In Figure 14.23 we show a variety of surfaces with boundary. On the left, S_1 is a genus 3 surface from which the interior of two disks have been removed. In the middle, S_2 is a disk with two overlapping strips attached to its boundary. On the right, S_3 is a subspace of $\mathbb{R}^3$ whose boundary (considering S_3 as a surface with boundary) is a trefoil knot.

$$S_1 \qquad\qquad S_2 \qquad\qquad S_3$$

FIGURE 14.23: Examples of surfaces with boundary.

In "surface with boundary" the meaning of "boundary" is distinct from its previous meaning, although the resulting sets could coincide. If a disk is considered as a subset of $\mathbb{R}^2$, then its boundary as a surface with boundary is the same as its boundary as a subset of $\mathbb{R}^2$. (See Figure 14.24.) However, in the case of the Möbius band embedded in $\mathbb{R}^3$, its boundary as a surface with boundary is a circle, whereas its boundary as a subset of $\mathbb{R}^3$ is the entire Möbius band. ·

The boundary of D as a manifold with boundary

The boundary of M as a manifold with boundary

The boundary of D as a subset of $\mathbf{R}^2$

The boundary of M as a subset of $\mathbf{R}^3$
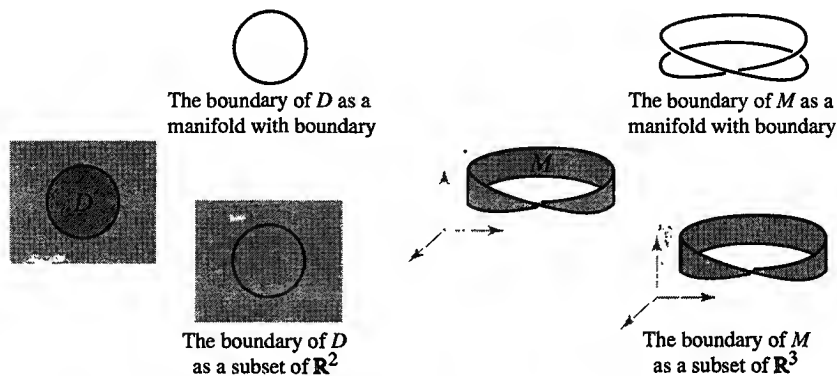
FIGURE 14.24: The boundary in a surface with boundary may or may not coincide with the boundary as a subset of a topological space.

We will accept the following facts about surfaces with boundary S. They are plausible, but to prove them requires preliminary results that would take us too far afield to pursue.

(i) The boundary of S is a 1-manifold.

(ii) If S is compact, then its boundary has finitely many components.

Much of what has already been addressed for compact surfaces (the existence of triangulations, the definition of orientability, the Euler characteristic, and so on) carries over to compact surfaces with boundary.

Let S be a compact surface with boundary. Each component of its boundary is a compact 1-manifold, and therefore by Theorem 14.3, each such component must be homeomorphic to a circle. For each such boundary circle, we can add a disk with its boundary glued to the circle. This process is called **capping off** the boundary of S. Then, S together with these glued-on disks is a compact surface S^* with no boundary. Therefore, every compact surface S with boundary is obtained by removing the interiors of finitely many disjoint disks from a compact surface S^* among those listed in Theorem 14.20.

If a surface with boundary S is embeddable in $\mathbb{R}^3$, it need not be the case that the corresponding capped-off surface S^* is as well. For example, the Möbius band is a surface with boundary that is embeddable in $\mathbb{R}^3$, but when we cap off a Möbius band we obtain a projective plane, a surface that is not embeddable in $\mathbb{R}^3$.

In order to identify a compact surface with boundary S, it is enough to calculate its Euler characteristic, determine whether or not it is orientable, and determine how many boundary components it has. Then we consider the capped off surface S^*. It is straightforward to see that the Euler characteristics of S and S^* are related by $\chi(S^*) = \chi(S) + n$, where n is the number of boundary components of S or, alternatively, the number of disks glued onto S to obtain S^*. Using the Euler characteristic of S^* and whether or not S^* is orientable, we can identify S^* by Theorem 14.20. Then we have recognized S to be the resultant surface S^* with the interiors of n disjoint disks removed.

EXAMPLE 14.6. In Figure 14.25 we show a surface with boundary, S, that we consider as a subspace of $\mathbb{R}^3$. The boundary of S is connected and forms the knot denoted 7_4. We would like to identify S.
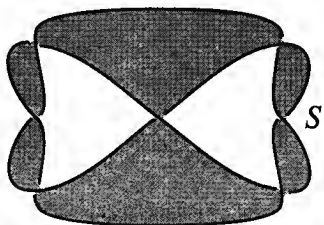
FIGURE 14.25· Which surface with boundary is this?

Note that every loop in S must pass through an even number of half twists. Therefore, it is not possible to embed a Möbius band in S, implying that S is orientable. Furthermore, the Euler characteristic of S is -1. (See Exercise 14.16.) Thus, if S^* is the surface obtained by capping off S, then S^* is orientable and $\chi(S^*) = \chi(S) + 1 = 0$. It follows that S^* is a torus, and S is the surface with boundary obtained by removing the interior of a single disk from a torus.

Exercises for Section 14.2

14.11. Prove Theorem 14.21: For every integer $n \geq 3$,

$$nP \cong \begin{cases} \frac{n-1}{2}T\#P & \text{if } n \text{ is odd,} \\ \frac{n-2}{2}T\#K & \text{if } n \text{ is even.} \end{cases}$$

14.12. Prove Theorem 14.24: For the compact surfaces nT and nP, the Euler characteristic is given by $\chi(nT) = 2 - 2n$ and $\chi(nP) = 2 - n$.

14.13. Consider the surface $T\#K\#K\#P\#P\#P$. Which of the surfaces listed in Theorem 14.20 is this? Which of the surfaces listed in Theorem 14.22 is this?

14.14. Consider all of the possible ways of creating a topological space from a square by gluing the edges in pairs. For each possibility, identify the space that results.

14.15. **(a)** Show that a hexagon with opposite edges identified straight across yields a torus.

(b) Determine what surface is obtained when opposite edges are identified straight across on an octagon.

(c) Determine what surface is obtained when opposite edges are identified straight across on a decagon.

(d) Determine what surface is obtained when opposite edges are identified straight across on a $2n$-gon, for $n \geq 6$.

14.16. For the surface with boundary S in Example 14.6, show that $\chi(S) = -1$.

14.17. Identify each of the surfaces with boundary shown in Figure 14.26.
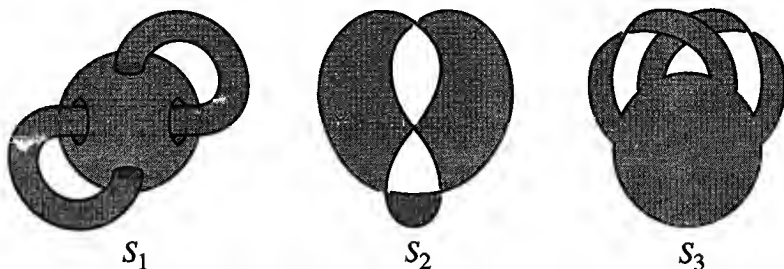
FIGURE 14.26: Identify each of these surfaces with boundary.

14.18. **(a)** Identify (without proof) all of the connected 1-manifolds with boundary.

(b) Identify (without proof) all of the compact 1-manifolds with boundary.

14.3   Three-Manifolds

In this section, we look at a variety of examples of 3-manifolds, keeping in mind that any one might be the spatial universe in which we reside.

EXAMPLE 14.7. The 3-sphere S^3. The 3-sphere is defined as the set of points at a unit distance from the origin in 4-space,

$$S^3 = \{(x, y, z, w) \mid x^2 + y^2 + z^2 + w^2 = 1\}.$$

It is a compact connected 3-manifold. (See Exercise 14.19.)

In Example 7.20, we discussed how S^3 can be regarded as the one-point compactification of 3-dimensional Euclidean space, $\mathbb{R}^3$. By adding just the single point to $\mathbb{R}^3$ and appropriately defining a topology on the result, we obtain a space homeomorphic to S^3.

Here we consider another way to picture S^3. We look at it in analogy to the 2-sphere, which can be thought of as an upper hemisphere glued to a lower hemisphere along an equator. Since the two hemispheres are each homeomorphic to the disk, we can think of the 2-sphere as coming from gluing the boundaries of two disks together. (See Figure 14.27.)
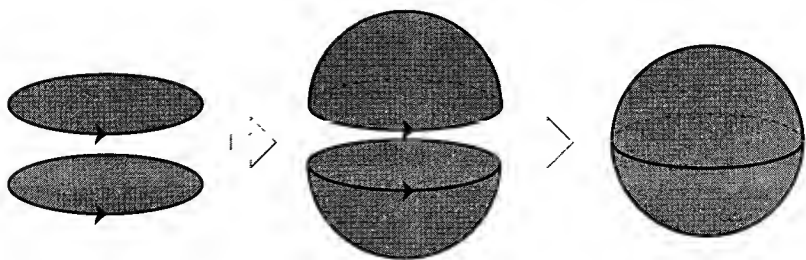
FIGURE 14.27: Gluing two disks together to obtain the 2-sphere.

We can similarly construct the 3-sphere by gluing two 3-balls together along their spherical boundary. But where are these two 3-balls in our initial description of the 3-sphere as the set of points at a unit distance from the origin in 4-space? Let

$$R = \{(x, y, z, w) \mid x^2 + y^2 + z^2 + w^2 = 1 \text{ and } w = 0\}.$$

In other words, R is the subset of S^3 obtained by setting $w = 0$. Notice that for those points, $x^2 + y^2 + z^2 = 1$, so R is a 2-sphere. The two sets

$$S_-^3 = \{(x, y, z, w) \mid x^2 + y^2 + z^2 + w^2 = 1 \text{ and } w \leq 0\} \text{ and}$$

$$S_+^3 = \{(x, y, z, w) \mid x^2 + y^2 + z^2 + w^2 = 1 \text{ and } w \geq 0\}$$

each have boundary R. Therefore S^3 is obtained by gluing S_-^3 and S_+^3 together along their spherical boundaries, and each of these subspaces of S^3 is homeomorphic to the 3-ball. (See Exercise 14.20.) Thus, in this way, we can see that S^3 results from gluing two 3-balls along their boundary.

We would like to have a simple means for generating examples of 3-manifolds. Just as surfaces can be triangulated, 3-manifolds can be decomposed into tetrahedra, any two of which can meet in a face, an edge, or a vertex, if at all. And just as we can glue the finite number of triangles in a triangulation of a compact connected surface together along a subset of the edge gluings to realize the surface as a polygon with pairs of edges to be identified, we can glue the finite number of tetrahedra in a triangulation of a compact, connected 3-manifold together along a subset of the face gluings to realize the 3-manifold as a polyhedron with pairs of faces on its exterior to be identified.

When discussing polyhedra, by the interior of a face we mean the set obtained by removing the perimeter edges and vertices from the face. Similarly, by interior of an edge, we mean the set obtained from the edge by removing the vertices on either end of it. These interiors correspond to the interior of the sets as manifolds with boundary, rather than the interior as subsets of the polyhedron.

EXAMPLE 14.8. The 3-torus. In Chapter 3 we defined the 3-torus as the product space $S^1 \times S^1 \times S^1$. The circle S^1 has a countable basis and is compact, connected, and Hausdorff. It follows that the 3-torus, being a product of circles,

has these properties as well. Each point in S^1 has a neighborhood that is homeomorphic to an open interval. Therefore each point in the 3-torus has a neighborhood that is homeomorphic to a product of three open intervals, and such a neighborhood is homeomorphic to the open 3-ball. Hence, the 3-torus is a compact, connected 3-manifold.

We can also obtain a 3-torus by gluing opposite faces straight across on a cube, as in Example 3.26. For similar constructions that we do hereafter, it is helpful to see in this representation how each point in the 3-torus has an open 3-ball neighborhood. A point in the interior of the cube has an open 3-ball neighborhood, and therefore so does the corresponding point in the 3-torus. (See Figure 14.28.) A point in the interior of a face has a half-ball neighborhood. That point is identified with a point in the opposite face that also has a half-ball neighborhood. These two half-ball neighborhoods glue together to form an open 3-ball neighborhood of the corresponding point in the 3-torus. A point in the interior of an edge has a quarter-ball neighborhood that glues together with the quarter-ball neighborhoods of three corresponding points in three other edges to form an open 3-ball neighborhood of the resulting point in the 3-torus. Finally, the eight points at the vertices have eighth-ball neighborhoods that also glue together to form an open 3-ball neighborhood about the resultant point in the 3-torus. Thus, every point in the 3-torus has an open 3-ball neighborhood.
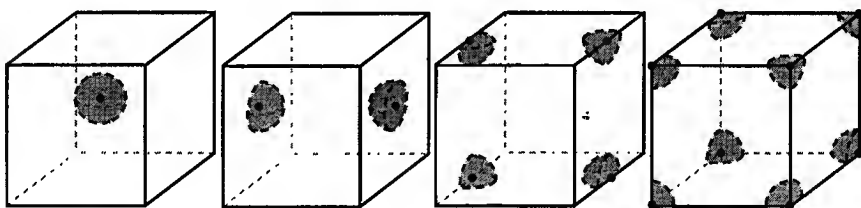
FIGURE 14.28: Each point in the 3-torus has an open 3-ball neighborhood.

Based on this example, it seems relatively straightforward to generate compact, connected 3-manifolds. Just take your favorite polyhedron, with an even number of n-gon faces for each n, and glue the faces together in pairs so that edges glue to edges and vertices glue to vertices. We refer to this process of constructing a quotient space as a **polyhedron gluing**.

We have to be careful, though, because the result of a polyhedron gluing might not be a 3-manifold. Given a polyhedron gluing, the points in the resulting quotient space that correspond to interior points of the polyhedron have open 3-ball neighborhoods, as needed. Also, the points that result from gluing points in the interior of the polyhedron's faces have open 3-ball neighborhoods, obtained from gluing two half-ball neighborhoods, as in the case of the 3-torus. We are not guaranteed however, that the points that arise from gluing the edges of the polyhedron, or from gluing the vertices of the polyhedron, have open 3-ball neighborhoods. In order to obtain a 3-manifold from a polyhedron gluing, there are additional conditions on the edge gluing and the vertex gluing that need to be satisfied. We discuss these conditions next.

To begin, we consider the edges. Given a polyhedron gluing, an equivalence relation is naturally defined on the edges such that two edges are equivalent if they are glued to each other. In each equivalence class of edges, all of the edges are glued together. Suppose we have an equivalence class of edges, $\{e_1, \ldots, e_p\}$. Let e^* denote the set in the quotient space that results from gluing together the interiors of these edges. We assume that the midpoints of the edges are glued together, and we refer to the resulting quotient-space point as the midpoint of e^*.

Each point in the interior of an edge in the polyhedron has a neighborhood that resembles a section of an orange, as illustrated in Figure 14.29. Under the polyhedron gluing, orange-section neighborhoods are glued together to form quotient-space neighborhoods of each point in e^*. We are interested in whether or not these quotient-space neighborhoods are homeomorphic to the open 3-ball.
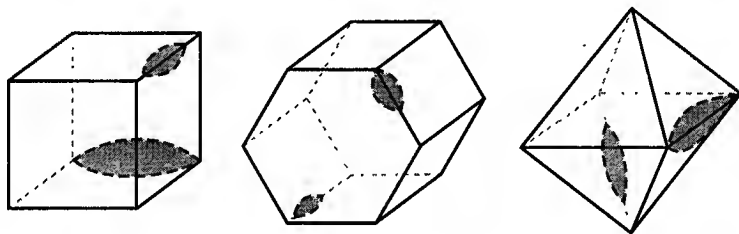
FIGURE 14.29: Points in the interior of each edge in the polyhedron have orange-section neighborhoods.

Assuming that the edges in our equivalence class are ordered properly, we can form a **face-edge sequence**

$$F_{[1,1]}, e_1, F_{[1,2]}, F_{[2,1]}, e_2, F_{[2,2]}, \ldots, F_{[p,1]}, e_p, F_{[p,2]},$$

such that edge e_i occurs on faces $F_{[i,1]}$ and $F_{[i,2]}$ for each $i = 1, \ldots, p$, face $F_{[i,2]}$ is glued to face $F_{[i+1,1]}$ for each $i = 1, \ldots, p-1$, and face $F_{[p,2]}$ is glued to face $F_{[1,1]}$. For each edge in our equivalence class, take an orange-section neighborhood that contains the entire interior of the edge. We can choose these neighborhoods so that they glue to each other to form a neighborhood of e^* in the quotient space. The face-edge sequence gives an ordering of the glued orange-section neighborhoods around e^*, and, since the last face in the sequence is glued to the first face, the orange-section neighborhoods glue together all of the way around e^*.

Now, if we put an arrow on e_1 to represent a direction, the other edges then inherit directions from e_1 as we glue the faces together in sequence. After we include e_p, all edges have a direction. The final face gluing then pairs e_p to e_1 so that either the arrows agree or they do not. In the former case, the orange-section neighborhoods glue together to form an open 3-ball neighborhood of each point on e^*. In the latter case, there are arbitrarily small neighborhoods of the midpoint of e^* that have projective-plane boundaries. (See Exercise 14.21.) While we do not prove it here, it can be shown that in the latter case no neighborhood of the midpoint of e^* is homeomorphic to the open 3-ball.

If the directions on e_p and e_1 agree, we say that the face-edge sequence is **direction preserving**; otherwise we say that it is **direction reversing**. It is straightforward to show that for an equivalence class of edges, if one face-edge sequence is direction preserving, then every face-edge sequence for the equivalence class is direction preserving (and therefore if one face-edge sequence for the equivalence class is direction reversing, then every face-edge sequence for the equivalence class is direction reversing). This leads to the following test:

The Edge Test: Given a quotient space obtained by a polyhedron gluing, for each equivalence class of edges, form a face-edge sequence and determine whether or not it is direction preserving. If each equivalence class of edges has a direction preserving face-edge sequence, then the quotient space passes the Edge Test; otherwise, the quotient space is not a 3-manifold.

For example, consider the polyhedron gluing for the 3-torus in Example 14.8. There are three equivalence classes of edges, and it is straightforward to see that each has a direction preserving face-edge sequence. Therefore this quotient space passes the Edge Test. Note that this is consistent with the observation in Example 14.8 that each point resulting from gluing the interiors of edges has an open 3-ball neighborhood obtained by gluing together quarter-ball neighborhoods of points in the edges of the cube.

EXAMPLE 14.9. Consider the gluing on the cube shown on the left in Figure 14.30. On the right in the figure, the edges numbered 1–4 form an equivalence class for this gluing. If we let T, B, F, B', L, and R, represent the top, bottom, front, back, left, and right faces of the cube, respectively, then $T, 1, F, B', 2, L, R, 3, B', F, 4, B$ is a face-edge sequence for the equivalence class. If we choose a direction for edge 1 as shown, then edges 2–4 inherit the depicted directions. The bottom face glues to the top face with a flip, and therefore the direction on edge 4 does not match the direction on edge 1 when those faces are glued. Thus, this quotient space fails the Edge Test and is not a 3-manifold.
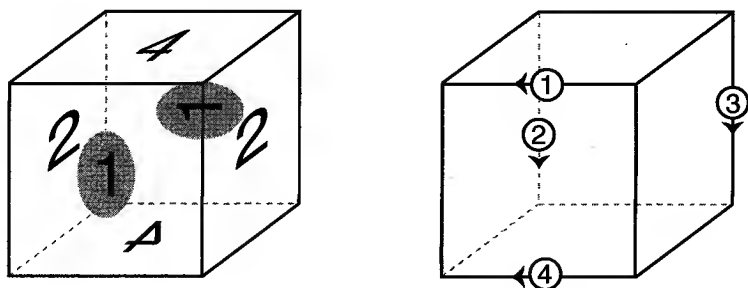
FIGURE 14.30: The equivalence class of edges $\{1, 2, 3, 4\}$ has a direction reversing face-edge sequence.

Next, let us consider the vertices in a polyhedron gluing. On the set of vertices, we define an equivalence relation such that two vertices are equivalent if they are glued together. For each equivalence class, all of the corresponding vertices are identified under the polyhedron gluing, resulting in a single point in the quotient space.

Suppose we have an equivalence class of vertices, $\{v_1, \ldots, v_n\}$. Let v^* denote the point in the quotient space that results from gluing together these vertices. When we cut through the polyhedron near a vertex v_i, as illustrated in Figure 14.31, each v_i is the apex of a pyramid that has faces lying in faces of the polyhedron and has a polygon base whose interior is in the interior of the polyhedron. (By pyramid, we mean a polyhedron formed by taking all of the line segments that connect an apex point to a base that is a polygon, not necessarily a square.)
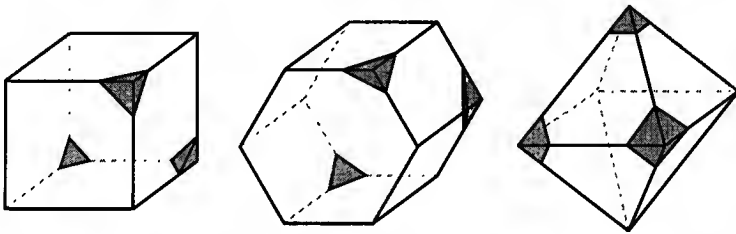
FIGURE 14.31: Each vertex of the polyhedron is the apex of a pyramid obtained by cutting the polyhedron near the vertex.

Now, take a collection of these pyramids, one for each v_i, such that when the polyhedron is glued together, the faces, edges, and vertices of the pyramids properly line up and are also glued together. Let B denote the subset of the quotient space resulting from gluing together the pyramids. The point v^* lies in the interior of B in the quotient space. We are interested in whether or not the interior of B is homeomorphic to the open 3-ball.

As a result of the way that the set B is constructed, the boundary of B is made up of the glued-together pyramid bases and consequently is a compact surface. We call this surface a **surrounding surface** for v^*. If this surface is a sphere, then as we shrink the pyramids toward the vertices v_i, the resulting surrounding surfaces essentially yield a set of concentric spheres closing in on v^*. It follows that B is homeomorphic to the 3-ball, and therefore v^* has an open 3-ball neighborhood. If the surrounding surface is not a sphere, it can be shown (but we do not present a proof) that no neighborhood of v^* is homeomorphic to the open 3-ball. This leads to the following test:

The Vertex Test: Given a quotient space obtained by a polyhedron gluing, if each equivalence class of vertices has a surrounding surface that is a sphere, then the quotient space passes the Vertex Test; otherwise, the quotient space is not a 3-manifold.

For example, consider the polyhedron gluing for the 3-torus in Example 14.8. There is one equivalence class of vertices, consisting of all eight vertices of the cube. Let v^* be the point in the quotient space resulting from

the glued-together vertices, and let S be a surrounding surface. Note that S is naturally triangulated by the pyramid bases since each base is a triangle.

We compute the Euler characteristic of S from this triangulation. Each vertex in the equivalence class contributes one pyramid base in forming S and therefore contributes one triangle to the triangulation. Thus, $F = 8$ for our Euler characteristic computation. There are $\frac{(3)(8)}{2}$ edges in the triangulation of the surrounding surface since each pyramid base contributes three edges, but these edges are glued together in pairs in forming S. Hence, $E = 12$. Finally, prior to our gluing the polyhedron, each of the cube's edges contains two vertices in the triangles that glue together to form S. Once the gluing is done, we have two triangulation vertices on each of the sets e^* that result from gluing together the cube's edges; that is, each equivalence class of cube edges contributes two vertices to the triangulation. We observed previously that there are three equivalence classes of edges; therefore $V = 6$. Consequently, the Euler characteristic of S is 2.

From our classification results for compact surfaces in the previous section, it follows that S is a sphere. Therefore, this quotient space passes the Vertex Test. Of course, this is consistent with the observation made in Example 14.8 that v^* has an open 3-ball neighborhood resulting from gluing together eighth-ball neighborhoods of each of the cube's vertices.

EXAMPLE 14.10. Consider the gluing on the octahedron shown on the left in Figure 14.32. The top vertex of the octahedron is glued to no other vertex and therefore is the only point in its equivalence class. Let v^* be the corresponding point in the quotient space. If we cut through the octahedron near the top vertex, we obtain a pyramid as illustrated in the middle of the figure. The front and back faces and the left and right faces of the pyramid are glued together, as in the octahedron gluing. The base of the pyramid is glued together on its edges to form a surrounding surface for v^*. As illustrated in the figure, this gluing amounts to identifying the opposite edges of a square in the usual way to obtain a torus. Thus, this quotient space fails the Vertex Test and is not a 3-manifold.
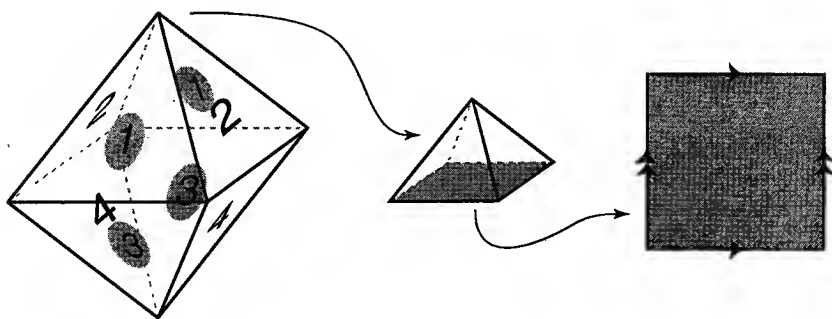
FIGURE 14.32: This octahedron gluing results in a point with a torus for a surrounding surface.

Combining our manifold-construction tests, we have the following theorem on polyhedron gluing:

THEOREM 14.27. *Let M be a quotient space resulting from a polyhedron gluing. Then M is a 3-manifold if and only if it satisfies both the Edge Test and the Vertex Test.*

> *Proof.* As already discussed, if the quotient space does not satisfy either the Edge Test or the Vertex Test, then M is not a 3-manifold.
>
> Thus suppose that the quotient space M satisfies both the Edge Test and the Vertex Test. That M is Hausdorff and has a countable basis follows in a manner similar to the case for the Klein bottle as a quotient space of a square, presented in Theorem 14.5. Small, open 3-ball neighborhoods, glued half-ball neighborhoods, glued orange-section neighborhoods, and glued neighborhoods of vertices can be used to define a countable basis and to separate points to show that M is Hausdorff.
>
> As previously discussed, the points in M that correspond to interior points of the polyhedron and the points in M that correspond to the interior of a face have open 3-ball neighborhoods. Since M satisfies the Edge Test, the points in M that result from gluing together the interiors of edges have open 3-ball neighborhoods obtained from gluing orange-section neighborhoods together. And, since M satisfies the Vertex Test, the points in M that result from gluing together vertices have 3-ball neighborhoods that come from taking the interior of glued-together pyramids.
>
> Therefore, if M satisfies the Edge Test and the Vertex Test, then M is a 3-manifold.   ∎

Let us consider some additional examples of 3-manifolds.

EXAMPLE 14.11. The quarter-twist manifold. This manifold is obtained from a cube by gluing two pairs of opposite faces straight across and by gluing the last pair by a one-quarter twist, as shown in Figure 14.33. There are three equivalence classes of edges, each containing four edges, as occurs for the 3-torus. Each of these equivalence classes has a direction preserving face-edge sequence. (See Exercise 14.22.)
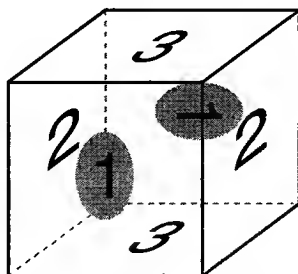
FIGURE 14.33: Gluing the faces of a cube to obtain the quarter-twist manifold.

All eight vertices in the cube are identified with each other, and, as in the 3-torus example, the Euler characteristic can be used to show that a surrounding surface of the corresponding point in the quotient space is a sphere. It follows that the quarter-twist manifold is a 3-manifold.

EXAMPLE 14.12. The sixth-twist manifold. This quotient space is defined on a hexagonal prism, as shown in Figure 14.34. Each square face is glued directly to the square face opposite it, and the front and back hexagonal faces are glued together with a one-sixth twist. The quotient space satisfies the Edge Test and the Vertex Test and therefore is a 3-manifold. (See Exercise 14.23.)
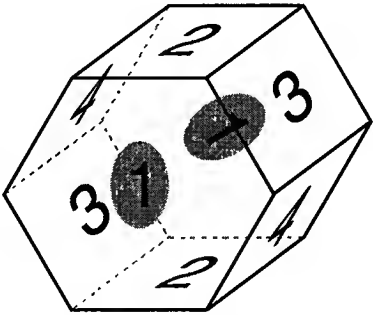
FIGURE 14.34: Gluing the faces of a hexagonal prism to obtain the sixth-twist manifold.

EXAMPLE 14.13. The Poincaré dodecahedral space. We identify each of the opposite pairs of faces on a dodecahedron by a one-tenth clockwise rotation, as in Figure 14.35. Note that a one-tenth clockwise twist identification from front to back is the same as a one-tenth clockwise twist identification from back to front. The resulting quotient space satisfies the Edge Test and the Vertex
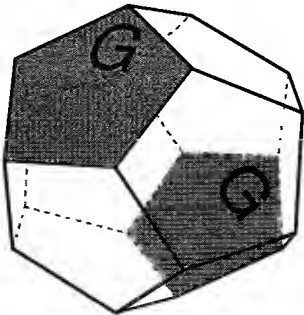
FIGURE 14.35: Gluing the faces of a dodecahedron to obtain the Poincaré dodecahedral space.

Test and therefore is a 3-manifold. (See Exercise 14.24.) The 30 edges are divided up into ten equivalence classes, each containing three edges. This will be relevant in the next section. The vertices are divided into five equivalence classes, each containing four vertices.

EXAMPLE 14.14. The Seifert–Weber dodecahedral space. In this case, we identify opposite faces on a dodecahedron by a three-tenths clockwise twist. (See Figure 14.36.) Here the 30 edges of the dodecahedron are divided into six equivalence classes of five edges each. All of the 20 vertices are identified with each other, and therefore there is just a single equivalence class of vertices. The quotient space is a 3-manifold. (See Exercise 14.25.)
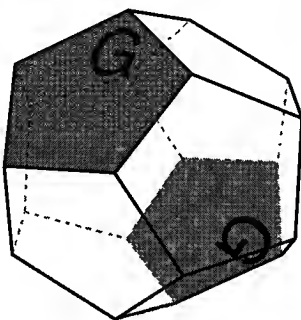
FIGURE 14.36: Gluing the faces of a dodecahedron to obtain the Seifert–Weber dodecahedral space.

As with 2-manifolds, we have a notion of orientability for 3-manifolds. Instead of nonorientability corresponding to the presence of a Möbius band, as in the case of a nonorientable 2-manifold, a 3-manifold is nonorientable if it contains a subspace known as a solid Klein bottle.

DEFINITION 14.28. *A solid Klein bottle is the quotient space obtained by identifying the opposite disks at the ends of a solid cylinder by reflection, as in Figure 14.37.*

FIGURE 14.37: A solid Klein bottle.

Notice that each concentric cylinder inside the solid cylinder becomes a Klein bottle after the gluing. So a solid Klein bottle consists of a set of concentric Klein bottles shrinking down to a core circle.

What would happen if we were inside a solid Klein bottle and we walked out one end, through the gluing, and came back to where we started?

Because of the way the gluing flips everything left to right, we would return with our hearts on the other side of our bodies. The faces of our watches would be reversed, and the hands on our watches would now move counterclockwise. Interestingly, we would not be able to tell the difference. To us, it would appear that our hearts were still on the left, and that our watches ran as they always did. But everyone that we left behind would appear reversed to us, as would we to them.

Could the universe contain such a solid Klein bottle? From a mathematical perspective, that is possible.

DEFINITION 14.29. *A 3-manifold is **nonorientable** if it contains a solid Klein bottle. If it does not, then the 3-manifold is **orientable**.*

EXAMPLE 14.15. Consider the quotient space obtained from a cube by identifying opposite faces as in Figure 14.38. We previously saw this space in Example 3.27. The front and back faces are identified straight across. The top and bottom faces are identified with a flip. The same is true for the left and right faces. The result of the gluing is a 3-manifold and is homeomorphic to $S^1 \times P$, the product of a circle and a projective plane. (See Exercise 3.38.)
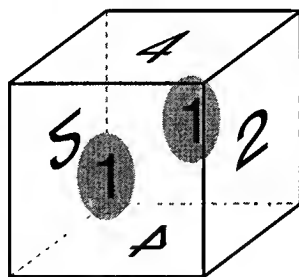
FIGURE 14.38: Identifying opposite faces of a cube to obtain a space homeomorphic to $S^1 \times P$.

This is a nonorientable manifold since it contains a solid Klein bottle, obtained by gluing the ends of a solid cylinder that runs from the center of the top face of the cube to the center of the bottom face.

One of the great, open questions in the field of 3-manifold theory was the Poincaré Conjecture. First posed by Henri Poincaré in 1904, it asked whether a 3-manifold that behaved enough like the 3-sphere must in fact be the 3-sphere.

In what sense must the manifold behave like the 3-sphere? Notice that if you take a loop in the 3-sphere (a continuous function $f : S^1 \to S^3$), then it can be shrunk down to a point. Since the 3-sphere is path connected and every loop in the 3-sphere can be shrunk to a point, the 3-sphere is said to be simply connected. (See Section 9.5.) In contrast, this is not the case with the 3-torus.

Under the polyhedron gluing that yields the 3-torus, a vertical line segment in the cube becomes a loop that cannot be shrunk to a point in the 3-torus.

The Poincaré Conjecture states that if a 3-manifold without boundary is compact and simply connected, then it must be homeomorphic to the 3-sphere. Poincaré attempted to prove this and failed. Many other attempts were made over the last century, and there were a number of announcements of proofs that were subsequently determined to be incorrect. In 2002, a Russian mathematician, Grigori Perelman, announced a proof that was judged correct after several years of scrutiny by mathematicians around the world. Not only did Perelman solve this celebrated, century-old problem, but his techniques have applications beyond the proof of the Poincaré Conjecture and promise to revolutionize research in three-dimensional topology. In 2006 Perelman was selected to receive the Fields medal for his work (but he declined the award).

Exercises for Section 14.3

14.19. Prove that the 3-sphere, S^3, is a 3-manifold.

14.20. Consider the spaces S^3_+ and S^3_- from Example 14.7. Define an explicit homeomorphism between each of these spaces and the 3-ball.

14.21. Suppose we have a polyhedron gluing and $\{e_1, \dots, e_p\}$ is an equivalence class of edges that has a direction reversing face-edge sequence. Let p be the point in the quotient space corresponding to the glued-together midpoints of the edges e_i. Show that in the quotient space there are arbitrarily small neighborhoods of p whose boundary is a projective plane.

14.22. (a) Show that the gluing in constructing the quarter-twist manifold in Example 14.11 satisfies the Edge Test and the Vertex Test.

    (b) Replace the quarter-twist gluing of the front and back faces of the cube by a half-twist gluing. Show that the resulting gluing satisfies the Edge Test and the Vertex Test. (The resulting manifold is called the **half-twist manifold**.)

14.23. (a) Consider the sixth-twist manifold construction in Example 14.12. On the hexagonal prism, label the edge classes and vertex classes according to the prescribed gluing. Show that the gluing satisfies the Edge Test and the Vertex Test.

    (b) Replace the one-sixth twist for gluing the front and back faces of the hexagonal prism with a one-third twist. Repeat part (a) for this prescribed gluing of the faces of the hexagonal prism. (The resulting manifold is called the **third-twist manifold**.)

14.24. Consider the gluing for the Poincaré dodecahedral space in Example 14.13.

    (a) Show that the gluing results in 10 equivalence classes of edges, each containing three edges, and five equivalence classes of vertices, each containing four vertices.

    (b) Show that the prescribed gluing satisfies the Edge Test and the Vertex Test.

14.25. Consider the gluing for the Seifert–Weber dodecahedral space in Example 14.14.

    (a) Show that the gluing results in six equivalence classes of edges, each containing five edges, and one equivalence class of vertices containing all the vertices.

    (b) Show that the prescribed gluing satisfies the Edge Test and the Vertex Test.

14.26. The **quaternionic manifold** is defined by identifying each pair of opposite faces in a cube with a one-quarter clockwise turn.

(a) On a diagram of a cube, label the faces to depict the gluing for this space.

(b) Show that the prescribed gluing satisfies the Edge Test and the Vertex Test.

14.27. Consider the gluing of faces of the cube that yields a space homeomorphic to $S^1 \times P$, as described in Example 14.15. On the cube, label the edge classes and vertex classes according to the prescribed gluing. Show that the gluing satisfies the Edge Test and the Vertex Test.

14.28. **(a)** Represent the product of a circle and a Klein bottle, $S^1 \times K$, as a quotient space obtained by gluing faces of a cube.

(b) On the cube, label the edge classes and vertex classes according to the prescribed gluing. Show that the gluing satisfies the Edge Test and the Vertex Test.

(c) Show that $S^1 \times K$ is nonorientable.

14.29. Glue the faces of a tetrahedron as shown in Figure 14.39. Is the result a 3-manifold?
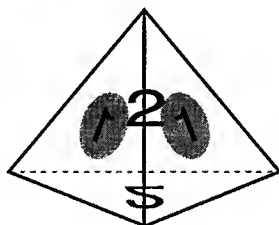
FIGURE 14.39: Does a 3-manifold result from this polyhedron gluing?

14.30. Consider the octahedron gluings described in (a) and (b) and illustrated in Figure 14.40. In each case determine whether or not the gluing results in a 3-manifold.

(a) The faces are glued together in pairs across the top and across the bottom, in each case with a one-third clockwise twist.

(b) The faces are glued together in opposite pairs through the center of the octahedron, in each case with a one-sixth counterclockwise twist.
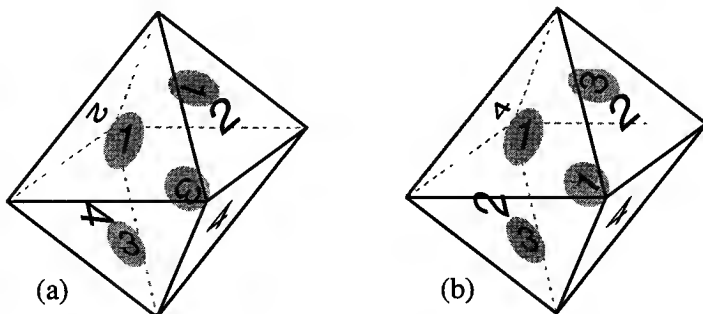
FIGURE 14.40: Does a 3-manifold result from either polyhedron gluing?

14.4   *The Geometry of the Universe*

Let us return to considering the spatial universe in which we live. As we noted, it seems to have the property that around every point there is a neighborhood that is homeomorphic to an open ball in $\mathbb{R}^3$. We assume that the spatial universe is Hausdorff and has a countable basis for its topology and therefore is a 3-manifold. But which 3-manifold is it?

We can use geometry to limit the possibilities. We will not be formal about what it means to have a geometry on a manifold, but it suffices to think of it as meaning that there is additional structure that allows the measurement of quantities such as distance, angles, and curvature. We are all familiar with Euclidean geometry on the line, the plane, and 3-space. As geometries, these spaces are denoted by E, E^2, and E^3, respectively.

The geometry of the universe seems to be isotropic and homogeneous. To be isotropic means that at each point the geometry appears to be the same in all directions around the point. There is no preferred identifiable direction. This is a property that Euclidean 3-space E^3 has, for instance. However, if we form a geometry by taking $S^2 \times E$, for example, that geometry has different behavior in different directions.

To be homogeneous means that locally the geometry of the space is the same. Given any two points in the space, there is an isometry (a distance-preserving homeomorphism) from a neighborhood of one point to a neighborhood of the other. For the purpose of understanding the overall, global geometry of the inverse, we ignore the problems that may occur around black holes; they have an isolated effect that we assume does not disturb the global geometry.

There are only three three-dimensional geometries that are both homogeneous and isotropic: the Euclidean, spherical, and hyperbolic geometries. Just as $\mathbb{R}^3$ is the model space for 3-manifolds, Euclidean 3-space E^3, the 3-sphere S^3, and a space known as hyperbolic 3-space are the models for geometric 3-manifolds. By a **Euclidean 3-manifold**, we mean a 3-manifold on which there is defined a geometry that locally behaves like the geometry in Euclidean 3-space. We similarly define a **spherical 3-manifold** and a **hyperbolic 3-manifold**. We expect the universe to be one of these three types of 3-manifolds, with a geometry that locally behaves just like the geometry in the corresponding model space.

Each of the model three-dimensional geometries has a curvature associated to it. In the case of Euclidean geometry, the curvature is 0. In the case of spherical geometry, the curvature is positive, and in the case of hyperbolic geometry, the curvature is negative.

Looking at analogs of these three geometries one dimension down, we have a plane with zero curvature, a sphere with positive curvature, and a saddle with negative curvature, as illustrated in Figure 14.41.

On a plane, which has zero curvature, the angles of a triangle add up to exactly 180 degrees. On a sphere, which has positive curvature, the angles of a triangle add up to more than 180 degrees. And, on a saddle, which has negative curvature, the angles of a triangle add up to strictly less than 180 degrees.
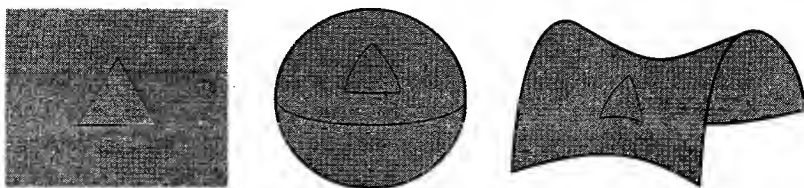
FIGURE 14.41: Geometries with zero curvature, positive curvature, and negative curvature.

Suppose we have a 3-manifold M obtained by gluing the faces of a polyhedron in Euclidean space. In order for M to be a Euclidean 3-manifold, there are some conditions that must be satisfied. First, the gluing between the faces must preserve distance on the faces, so that the local geometry is Euclidean geometry at each point that results from gluing together two faces. Then, given that the faces are glued together preserving distance, when we glue orange-section neighborhoods together to obtain a neighborhood of a point in M, the angles between the orange-section faces that glue together must add up to 360 degrees. In other words, it is necessary that each point in M that results from gluing together edges has a full 360 degrees of Euclidean angle around it. Finally, it must also be the case that the manifold inherits the correct geometry around the glued vertices. Small neighborhoods of polyhedron vertices that glue to each other must fit together so that the local geometry is Euclidean geometry in a neighborhood of the resultant point in the manifold. In all of the following examples, once the edge conditions are satisfied, the glued-together vertices inherit the required neighborhoods.

We can also glue together the analogs of polyhedra in spherical space, and we can do the same in hyperbolic space. In both settings, in order for the result to be a 3-manifold with the corresponding geometry, we need to satisfy conditions like those just discussed for Euclidean 3-manifolds.

EXAMPLE 14.16. Let M be a 3-manifold that results from gluing the faces of a cube in Euclidean space. At each edge in a cube, two faces meet with a 90 degree angle between them. In order for M to be a Euclidean 3-manifold, it is necessary to have four edges in each equivalence class of edges, so that we then have the requisite 360 degrees in neighborhoods of points in M that result from gluing together edges in the cube. This occurs for the 3-torus, the quarter-twist manifold, and the half-twist manifold introduced in Exercise 14.22. Therefore each of these is a Euclidean 3-manifold.

EXAMPLE 14.17. Consider the sixth-twist manifold from Example 14.12. It is formed by gluing the faces of a hexagonal prism in Euclidean 3-space. Let e be an edge at the intersection of two square faces in the prism. The angle between the faces is 120 degrees, and e glues together with two other edges that are each at the intersection of two square faces. Therefore, in the manifold each point that results from gluing these edges has a total angle around it of

360 degrees. Now, let e be an angle at the intersection of a hexagonal face and a square face. In this case, the angle between the faces is 90 degrees, and e glues together with three other edges that are each at the intersection of a hexagonal face and a square face. Here too, in the manifold each point that results from gluing these edges has a total angle around it of 360 degrees. It follows that the sixth-twist manifold is a Euclidean 3-manifold.

EXAMPLE 14.18. The angles between the faces of a Euclidean dodecahedron measure $\cos^{-1}(-\frac{1}{5}\sqrt{5}) \approx 116.6$ degrees. For the gluing on the dodecahedron that yields the Poincaré dodecahedral space, we indicated that the edges are grouped into 10 equivalence classes of three edges each. This means that around the corresponding points in the manifold the total angle is shy of the needed 360 degrees. The resulting manifold is not a Euclidean 3-manifold.

To resolve this situation, we realize the dodecahedron in spherical geometry. Placing a dodecahedron in spherical geometry increases the angles at its edges. In fact, for any fixed value of θ between $\cos^{-1}(-\frac{1}{5}\sqrt{5})$ and 180 degrees, there is a dodecahedron in the 3-sphere having θ as the angle between each pair of intersecting faces. In particular, a regular dodecahedron exists in the 3-sphere with angles of 120 degrees between the faces at each of the edges. We can glue together such a dodecahedron to obtain the Poincaré dodecahedral space, and in so doing we satisfy the requirement of having 360 degrees around each point in each glued-together edge. Therefore, the Poincaré dodecahedral space is a spherical 3-manifold.

For the gluing on the dodecahedron that yields the Seifert–Weber dodecahedral space, the edges are grouped into six equivalence classes of five edges each. Hence, we need to have an angle of 72 degrees between the dodecahedron faces at each edge. Here too, we cannot realize the resulting manifold as a Euclidean 3-manifold. However, given any angle θ between 60 and $\cos^{-1}(-\frac{1}{5}\sqrt{5})$, in hyperbolic space we can construct a dodecahedron such that θ is the angle between each pair of intersecting faces. In particular, a regular dodecahedron with angles of 72 degrees between faces exists in hyperbolic space. Hence, the Seifert–Weber dodecahedral space is a hyperbolic 3-manifold.

Data from the Wilkinson Microwave Anisotropy Probe, launched by the National Aeronautics and Space Administration (NASA) in 2001, has provided cosmologists with some evidence that the curvature of the universe is very close to 0. This either means that we live in a Euclidean universe, or means that we live in a spherical or hyperbolic universe with low curvature. If we live in a Euclidean universe with curvature 0, the geometry that we experience in the universe, even on the largest scale, is Euclidean geometry, the one we have grown up studying. So a giant triangle with vertices formed by distant galaxies would have the property that the sum of its angles is 180 degrees. If we assume that the universe is Euclidean, does this help us determine which manifold is the universe? Which 3-manifolds have a Euclidean geometry? Remarkably, there are only 18 possibilities, a result first proved in 1934 by mathematical crystallographer Werner Nowacki (1909–1988) in [Now].

THEOREM 14.30. *There are exactly eighteen Euclidean 3-manifolds, six compact and orientable, four compact and nonorientable, four noncompact and orientable, and four noncompact and nonorientable.*

This theorem is an amazing result, telling us that the possibilities for Euclidean 3-manifolds are extremely limited. In contrast, there are infinitely many spherical 3-manifolds and infinitely many hyperbolic 3-manifolds.

If we assume further that the universe is compact, the number of possibilities drops to 10. Four of the 10 are nonorientable. As we indicated in the previous section, there are some unusual consequences of traveling through a solid Klein bottle in a nonorientable 3-manifold. As fascinating as such a possibility is, there are physics aspects that make it highly unlikely. Thus, what remains for consideration are six compact and orientable Euclidean 3-manifolds.

Of these six manifolds, we have already seen the 3-torus, the quarter-twist manifold, the sixth-twist manifold, the half-twist manifold, and the third-twist manifold, in the examples and exercises in Section 14.3. The remaining possibility, the Handtche–Wendt manifold, is introduced in Exercise 14.32.

Exercises for Section 14.4

14.31. Show that the third-twist manifold of Exercise 14.23 can be realized as a Euclidean manifold.

14.32. Define the Handtche–Wendt manifold by gluing together the faces on a pair of cubes as shown in Figure 14.42.
   (a) Show that the prescribed gluing satisfies the Edge Test and the Vertex Test and therefore the result is a 3-manifold.
   (b) Show that the Handtche–Wendt manifold is a Euclidean 3-manifold.
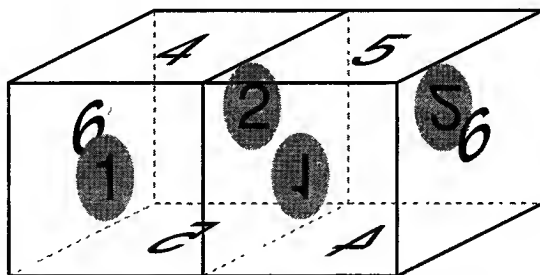
FIGURE 14.42: Gluing faces on a pair of cubes to obtain the Handtche–Wendt manifold.

14.33. The quaternionic manifold, introduced in Exercise 14.26, has one of Euclidean, spherical, or hyperbolic geometry. Which one is it?

14.34. Show how each of the first five compact, orientable Euclidean 3-manifolds is a so-called torus bundle. That is to say, there exists a torus embedded in each manifold such that when we remove a neighborhood of that torus, the result is homeomorphic to $T \times [0, 1]$, the product of a torus and an interval.

14.35. A parallelepiped is defined by any three linearly independent vectors in 3-space. For each of the three compact, orientable Euclidean 3-manifolds that can be obtained by identifying opposite faces of the cube, determine what restrictions on the shape of a parallelepiped are necessary for the manifold to be realized by identifying opposite faces of the parallelepiped.

14.36. Find examples of three different noncompact and connected Euclidean 3-manifolds.

14.37. (a) Show that $K \times S^1$, the product of a Klein bottle and a circle, can be realized as a Euclidean 3-manifold.

(b) Give an example of another compact, connected and nonorientable Euclidean 3-manifold.

14.5   *Determining which Manifold is the Universe*

Now that we have a list of some candidate manifolds for the spatial universe, how do we tell which is the right one? In this section, we discuss two methods that cosmologists are using to determine the shape of the universe.

Cosmic Crystallography

French cosmologists Marc Lachièze-Rey, Roland Lehoucq, and Jean-Pierre Luminet proposed this method in their 1996 paper [Leh]. Imagine we live in a 3-torus, formed from gluing together the faces of a cube that has an edge length of 10 million light-years. Further, for now, imagine that we can see everything that is in the universe at this instant (we will adjust this physically impossible assumption shortly). When we gaze out in each of the directions perpendicular to the cube's faces, as illustrated in Figure 14.43, our galaxy is in our line of sight, 10 million light-years away. Extending our line of sight further, we see our galaxy again 20 million light-years away, and then 30 million light-years away, and so on. Therefore, our galaxy forms a lattice pattern as we look out into the universe in all directions. Not only that, so does every galaxy we see within the glued-together cube. While galaxies may be scattered randomly within the cube, if we consider every galaxy image that we see as we gaze out, there are lots of pairs of images that are 10 million light-years apart, and there are lots of pairs of images that are 20 million light-years apart, and so forth.
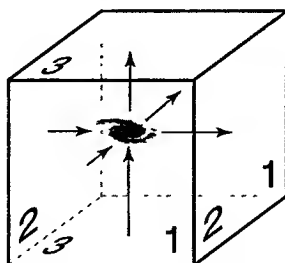
FIGURE 14.43: In a 3-torus universe, our line of sight brings us back to our galaxy.

In this 3-torus universe, it is not only at integer multiples of 10 million light-years that we see our galaxy. If we look in a direction parallel to a diagonal in one of the cube's faces, then we see our galaxy $\sqrt{2} \times 10$ million light-years away and at integer multiples of this distance. Also, if we look in a direction parallel to one of the cube's diagonals, then we see our galaxy $\sqrt{3} \times 10$ million light-years away and at integer multiples of this distance as well. There are other distances, too, at which we see our galaxy, depending on our viewing direction.

By this reasoning, if we live in this 3-torus universe, and we measure the distance between every pair of galaxies that we see, then in the overall distribution of distances, we expect to have spikes as illustrated on the left in Figure 14.44. In contrast, if our universe is Euclidean 3-space, we might expect to see a somewhat random distribution of distances, centered around some average distance, as shown on the right in Figure 14.44.
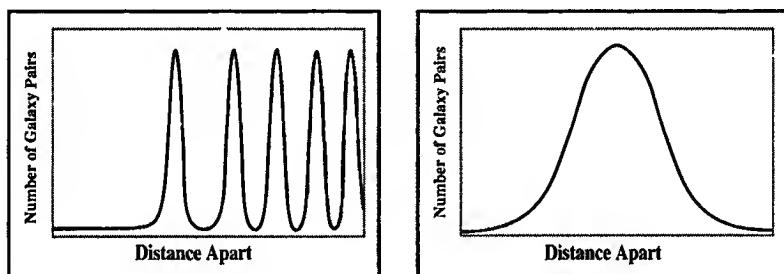
FIGURE 14.44: Hypothetical distributions of distances between pairs of galaxies in the 3-torus and in 3-space.

In theory, this sounds great; however, we have to take into account the physical reality. It takes time for the light from galaxies to reach us, and galaxies are in motion, so two different images of a particular galaxy will not necessarily have a distance between them exactly equal to one of the previously-mentioned theoretical values (10 million light-years, $\sqrt{2} \times 10$ million light-years, $\sqrt{3} \times 10$ million light-years, and so on). But still, we can expect to see distances clustered around these values. Furthermore, galaxies are hard to spot when they are a great distance away, so they are not the best structures to be observing.

An alternative that cosmologists are using is measuring the distances between superclusters of galaxies. The hope is to use catalogs of superclusters to identify the distances between pairs of superclusters, and then to use patterns in the distance data to try to identify the manifold corresponding to the shape of the universe. With current supercluster catalogs, no patterns have yet been discovered. But in the near future, more complete catalogs will become available, which may provide patterns that help reveal the manifold in which we live.

The Cosmic Microwave Background Radiation

In 1965, Arno Penzias and Robert Wilson of Bell Lab found an unexpected, constant background radio noise when studying radio emissions from the Milky Way. The scientists eventually realized that the noise they were receiving was essentially constant wavelength radiation coming from outside the galaxy in every direction in the sky. This radiation is now known as the cosmic microwave background (CMB). Their discovery turned out to have very important implications, as this radiation proved to be the oldest light in the universe. The light of the CMB marks the first time after the big bang that the universe had cooled enough for radiation to escape from the hot cosmic soup. About 300,000 years after the big bang, the temperature of the universe was sufficiently cool to allow electrons and protons to combine to form the first atoms. This also meant that radiation, which had previously been constantly scattered by the free charged particles, was now able to travel at the speed of light through the still expanding universe. This escaped radiation is the cosmic microwave background that we now observe as a relatively constant temperature (about 2.73 Kelvin) radio source in every direction. The temperature of the CMB varies to one part in 10 thousand. The small variations that exist in the CMB show us the slight density differences present in the early universe that eventually led to the clumping of matter into galaxies, stars, and planets through the force of gravity.

The CMB radiation appears to emanate from a large sphere, called the surface of last scattering. Data from NASA's Cosmic Background Explorer (COBE) satellite, launched in 1989, enabled the construction of the first temperature map of the CMB. NASA's Wilkinson Microwave Anisotropy Probe (WMAP), launched in 2001, has provided CMB temperature data of significantly higher resolution than COBE. In Figure 14.45, we show a temperature map of the CMB derived from WMAP data. The darker shaded areas are cooler and the lighter shaded areas are warmer, but, as already indicated, the temperature difference on either side of the 2.73 Kelvin average is only about 0.0001 K.
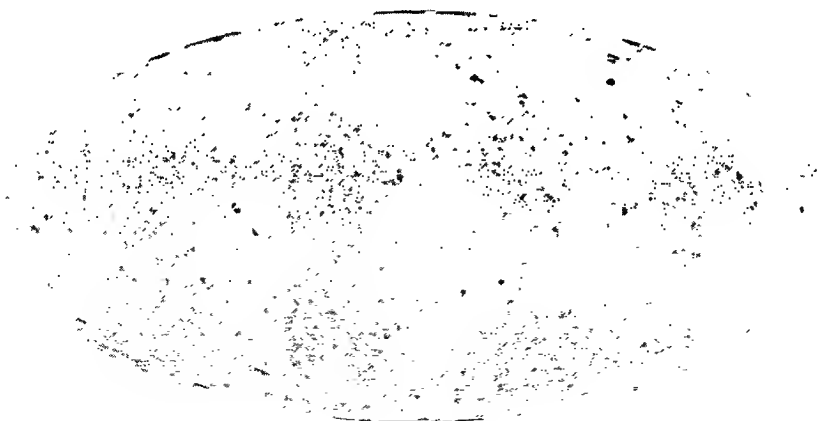
FIGURE 14.45· A temperature map of the CMB. (Courtesy of the NASA/WMAP Science Team)

Cosmologists Neil Cornish, David Spergel, and Glen Starkman, and mathematician Jeffrey Weeks have proposed an approach, called "circles in the sky," for determining the topology of the universe from the CMB temperature profile. (See [Cor1], [Cor2], and [Wee1].) Imagine that the universe is a 3-torus formed by gluing together opposite faces of a cube. Furthermore, imagine that the surface of last scattering expands to the point that its diameter exceeds the edge length of the cube, as illustrated in Figure 14.46. When the sphere intersects opposite faces in the cube, it causes the sphere to intersect itself since the opposite faces are glued together. As shown in the figure, these self-intersections result in a collection of circles on the sphere, and when we examine the sphere from the inside, each intersection circle can be seen in two different locations on the sphere.
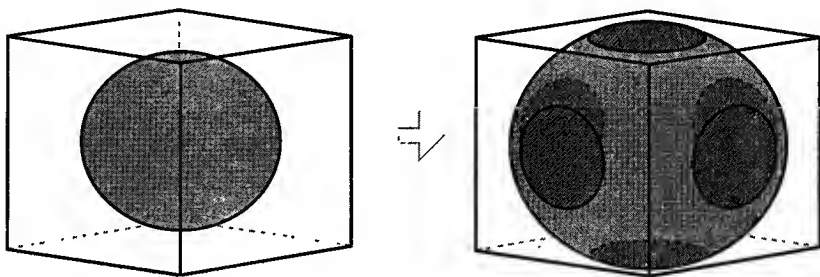
FIGURE 14.46: Circles in the sky result from the sphere intersecting itself.

The "circles in the sky" idea is to examine the CMB temperature map to find pairs of circles around which the temperature distributions are the same. Each such pair of circles could represent an intersection of the surface of last scattering with itself. The hope is that the arrangement of intersection circles on the surface of last scattering will reveal patterns that can only occur within one of the manifolds that is a candidate for the topological shape of the universe. In this way, we may be able to identify the manifold in which we live.

Exercises for Section 14.5

14.38. If we are in a 2-dimensional universe that is a torus, constructed from a rectangle with edge lengths 2 and 4, where do we see peaks in the distribution of intergalactic supercluster distances? If, instead, the edge of length 2 on the rectangle is glued to its opposite edge with a flip, so that the resultant space is a Klein bottle, where do we see the peaks?

14.39. How do the peaks in the distribution of intergalactic supercluster distances compare between the case where we glue opposite faces of a $3 \times 3 \times 4$ rectangular solid to obtain a 3-torus and the case where we glue the faces to obtain a half-twist manifold, where the twisted face is the one of dimensions 3×3. And how do those situations compare with the case where we glue the faces to obtain a quarter-twist manifold, where the twisted face is the one of dimensions 3×3?

14.40. Suppose that the universe is a cube with opposite pairs of faces glued to yield either the 3-torus, the quarter-twist manifold, or the half-twist manifold. Explain how we could use circles in the sky to distinguish between these possibilities, indicating the differences between the circles' appearance in each case.

Additional Readings

Here we list some additional readings related to the topics introduced in the text. The topics are listed in the order in which they first appear in the text. There is a wealth of literature on many of these topics; we primarily chose readings that provide a good starting point or a good overview.

History of Topology: The compilation [Jam] consists of 40 papers that address different topics in the history of topology. The book [Man] is an exposition of the role that concepts and problems in analysis played in the mathematical developments leading to the axiomatic foundation of topology.

Digital Topology: Along with [Kon], mentioned in the text, other papers that initially introduced topological spaces for modeling in digital image processing include [Kha1], [Kha2], and [Kov].

Phenotype Spaces: Besides the paper [Fon1], mentioned in the text, other papers that address phenotype spaces and continuity in evolution include [Cup1], [Cup2], [Fon2], [Sta1], and [Sta2].

Spatial Relations in Geographic Information Systems: There are many papers in the geographic information systems literature addressing topological spatial relations. The papers [Sch], [Sha], and [The] are a sample of works that are based on the model that is discussed in the text and that was presented originally in [Ege]. An alternative topological spatial relation model is presented ..in [Ren].

Topology and Physics: The texts [Mon], [NasC2], and [Shv], at three different levels of mathematical sophistication, introduce and discuss topics in topology and their role in physics. The paper [NasC1] presents a history of the connections between Physics and Topology from the mid-nineteenth century to the present.

The Forward Kinematics Map: The paper [Bak1] is a relatively nontechnical introduction to topological properties of the forward kinematics map. Other papers that provide an introductory perspective on this topic include [Bak2], [Dem], and [Got].

Error-Correcting Codes: The texts [Bay] and [Ple] present basic introductions to the mathematics of error-correcting codes.

Levenshtein Metric: The book [San] contains a series of papers about the Levenshtein metric; it includes applications to DNA sequence analysis, speech recognition, error correction, and bird-song analysis.

Automated Guided Vehicles: The text [Lat] presents a broad introduction to the mathematics of robot motion planning. Along with [Abr2], mentioned in the text, [Abr1], [Ghr1], and [Ghr2] are a few recent papers that examine the structure of configuration spaces for automated guided vehicles.

Dynamical Systems and Chaos: The texts [DevR], [All], and [Str] provide introductions to various aspects of the theory and applications of dynamical systems and chaos. The text [Gil] gives a more advanced treatment of dynamical systems theory from a topological perspective. A nontechnical overview of the development of chaos theory from the 1960s into the 1980s can be found in [Gle].

Economics Applications and Game Theory: The book [Von], by John Von Neumann and Oskar Morgenstern, is a foundational work on the mathematics of game theory and economic models. Kakutani's generalization of the Brouwer Fixed Point Theorem was originally published in [Kak], and Nash's equilibrium-existence theorem was first announced in [NasJ]. Recent texts covering fixed point theory and applications to economics and game theory include [Bor] and [Fra].

Knot Theory: The Reidemeister moves were presented by Kurt Reidemeister in [Rei1]; he later published a book on knot theory, [Rei2]. Vaughan Jones introduced his polynomial invariant in [Jon].

The books [Ada1], [Cro], and [Liv] are nice follow-ups to our material on knot theory; they expand on each of the topics that we present in Chapter 12, and they introduce a number of other topics in knot theory.

Knot Theory Applications: The text [Sum] contains six separate papers addressing applications of topology and knot theory in biology, chemistry, and physics. The paper [Die] presents the original announcement of the synthesis of a knotted molecule. The book [Fla] addresses a number of topics related to knots and embeddings of graphs, focusing on their relationship to the structure of molecules.

Topological Graph Theory: The book [Gro] is a nice introduction to topological graph theory, beginning with basic ideas in graph theory and covering many topics related to the topology of graphs.

Graph Applications in Chemistry: The text [DevJ] is a collection of papers addressing the use of topological descriptors in chemistry. The text [Tri] is a broad introduction to chemical graph theory, and the text [Bon] is a collection of nine separate articles on different aspects of this topic. Also, see the text [Fla] mentioned under Knot Theory Applications.

Graph Applications in Electronic Circuit Design: The publication [Roh] is a special issue of the IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems; it is devoted to routing in microelectronics, and it consists of 11 papers, most of which employ topological or graph-theoretic ideas. The papers [Agg], [Bha], and [Fou] each address applications of crossing number or thickness to electronic circuit design or construction.

Manifolds and Cosmology: The article [Ada2] is a nontechnical introduction to manifolds and their application to determining the shape of the universe. The paper [Thu2] presents a brief introduction to a variety of mathematical ideas in the study of manifolds. The book [Wee2] is an excellent introduction to the geometry and topology of surfaces and 3-manifolds, while [Thu1] is a more advanced text that focuses on 3-manifolds.

.

References

[Abb] Abbott, E.A., *Flatland: A Romance of Many Dimensions*, London: Seeley & Co., Ltd., 1884.

[Abr1] Abrams, A., "Configuration Spaces of Colored Graphs" *Geometriae Dedicata* **92** (2002), 185–194.

[Abr2] Abrams, A., and Ghrist, R., "Finding topology in a factory: configuration spaces" *American Mathematical Monthly* **109** (2002), no. 2, 140–150.

[Ada1] Adams, C., *The Knot Book: An Elementary Introduction to the Mathematical Theory of Knots*, Providence, R.I.: American Mathematical Society, 2004.

[Ada2] Adams, C., and Shapiro, J., "The Shape of the Universe: Ten Possibilities" *American Scientist* **89** (2001), no. 5, 443–453.

[Agg] Aggarwal1, A., Klawe, M., and Shor, P., "Multilayer grid embeddings for VLSI" *Algorithmica* **6** (1991), no. 1, 129–151.

[Ale] Alexander, J.W., "Topological invariants of knots and links" *Transactions of the American Mathematical Society* **30** (1928), 275–306.

[All] Alligood, K.T., Sauer, T.D., and Yorke, J.A., *Chaos–An Introduction to Dynamical Systems*, New York: Springer–Verlag, 1996.

[Bak1] Baker, D.R., "Some Topological Problems in Robotics" *The Mathematical Intelligencer* **12** (1990), 66–76.

[Bak2] Baker, D.R., and Wampler, C.W., "On the inverse kinematics of redundant manipulators" *International Journal of Robotics Research* **7** (1988), no. 2, 3–21.

[Ban] Banks, J., Brooks, J., Cairns, G., Davis, G., and Stacey, P., "On Devaney's Definition of Chaos" *American Mathematical Monthly* **99** (1992), 332–334.

[Bay] Baylis, J., *Error-Correcting Codes: A Mathematical Introduction*, London: Chapman & Hall Ltd., 1998.

[Bha] Bhatt, S.N., and Leighton, F.T., "A Framework for Solving VLSI Graph Layout Problems" *Journal of Computer and System Sciences* **28** (1984), 300–343.

[Blo] Bloch, E.D., *Proofs and Fundamentals: A First Course in Abstract Mathematics*, Boston: Birkhäuser, 2000.

[Bon] Bonchev, D., and Rouvray, D.H., eds., *Chemical Graph Theory: Introduction and Fundamentals*, New York: Abacus Press, 1991.

[Boo] Boone, W.W., Haken, W., and Poenaru, V., "On recursively unsolvable problems in topology and their classification" In: Schmidt, H.A., Schütte, K., and Thiele, H.J., eds., *Contributions to Mathematical Logic*, Amsterdam: North-Holland, 1968.

[Bor] Border, K.C., *Fixed Point Theorems with Applications to Economics and Game Theory*, Cambridge: Cambridge University Press, 1985.

[Chr] Christenson, C.O., and Voxman, W.L., *Aspects of Topology*, New York: Marcel Dekker Inc., 1977.

[Con] Connelly, R., and Demaine, E.D., "Geometry and Topology of Polygonal Linkages" In: Goodman, J.E., and O'Rourke, J., eds., *CRC Handbook of Discrete and Computational Geometry*, Second Edition, Boca Raton, Fl.: Chapman & Hall/CRC, 2004, 197–218.

[Cor1] Cornish, N.J., Spergel, D.N., and Starkman, G.D., "Circles in the sky: finding topology with the microwave background radiation" *Classical and Quantum Gravity* **15** (1998), 2657–2670.

[Cor2] Cornish, N.J., and Weeks, J.R., "Measuring the shape of the universe" *Notices of the American Mathematical Society* **45** (1998), no. 11, 1463–1471.

[Cro] Cromwell, P.R., *Knots and Links*, Cambridge: Cambridge University Press, 2004.

[Cup1] Cupal, J., Kopp, S., and Stadler, P.F., "RNA Shape Space Topology" *Alife* **6** (2000), 3–23.

[Cup2] Cupal, J., Schuster, P., and Stadler, P.F., "Topology in Phenotype Space" *Computer Science in Biology*, GCB'99 Proceedings, University Bielefeld, Hannover, 1999, 9–15.

[Dem] DeMers, D., and Kreutz-Delgado, K., "Learning Global Properties of Nonredundant Kinematic Mappings" *International Journal of Robotics Research* **17** (1998), no. 5, 547–560.

[DevR] Devaney, R.L., *An Introduction to Chaotic Dynamical Systems, 2nd Edition*, Boulder, Co.: Westview Press, 2003.

[DevJ] Devillers, J., and Balaban, A.T., eds., *Topological Indices and Related Descriptors in QSAR and QSPAR*, Amsterdam: Gordon and Breach Science Publishers, 1999.

[Die] Dietrich-Buchecker, C.O., and Sauvage, J.P., "A synthetic molecular trefoil knot" *Angewandte Chemie International Edition* **28** (1989), no. 2, 189–192.

[Dug] Dugundji, J., *Topology*, Boston: Allyn and Bacon, 1966.

[Ege] Egenhofer, E., and Franzosa, R., "Point-Set Topological Spatial Relations" *International Journal of Geographical Information Systems* **5** (1991), no. 2, 161–174.

[Far] Farber, M., "Collision Free Motion Planning on Graphs" In: Erdmann, M., Hsu, D., Overmars, M., and van der Stappen, A.F., eds., *Algorithmic Foundations of Robotics VI*, New York: Springer–Verlag, 2005.

[Fla] Flapan, E., `When Topology Meets Chemistry: A Topological Look at Molecular Chirality*, Washington D.C.: Mathematical Association of America, Cambridge: Cambridge University Press, 2000.

[Fon1] Fontana, W., "The Topology of the Possible" In: Wimmer, A., and Kössler, R., eds., *Understanding Change: Models, Methodologies and Metaphors*, New York: Palgrave Macmillan, 2006.

[Fon2] Fontana, W., and Schuster, P., "Continuity in Evolution: On the Nature of Transitions" *Science* **280** (1998), 1451–1455.

[Fou] Foulds, L.R., Perara, S.M., and Robinson, D.F., "Network Layout Procedure for Printed Circuit Design" *Computer Aided Design* **10** (1978), 441–451.

[Fra] Franklin, J., *Methods of Mathematical Economics*, New York: Springer–Verlag, 1980.

[Ghr1] Ghrist, R., "Configuration spaces and braid groups on graphs in robotics" In: Gilman, J., Menasco, W.W., and Lin, X.-S., eds., *Knots, Braids, and Mapping Class Groups—Papers Dedicated to Joan S. Birman*, Providence, R.I.: American Mathematical Society, 2001, 29–40.

[Ghr2] Ghrist, R., and Koditschek, D., "Safe, cooperative robot dynamics on graphs" *SIAM Journal of Control and Optimization* **40** (2002), no. 5, 1556–1575.

[Gil] Gilmore, R., and Lefranc, M., *The Topology of Chaos: Alice in Stretch and Squeezeland*, New York: John Wiley & Sons, Inc., 2002.

[Gle] Gleick, J., *Chaos: Making a New Science*, New York: Viking Penguin Inc., 1987.

[Got] Gottlieb, D.H., "Topology and the Robot Arm" *Acta Applicandae Mathematicae* **11** (1988), 111–121.

[Gro] Gross, J.L., and Tucker, T.W., *Topological Graph Theory*, New York: Wiley, 1987.

[Hau] Hausdorff, F., *Grundzüge der Mengenlehre*, Liepzig: Veit and Co., 1914.

[Hil] Hill, T.P., "Mathematical Devices for Getting a Fair Share" *American Scientist* **88** (2000), no. 4, 325–331

[Hum] Hummel, K.E., *Introductory Concepts for Abstract Mathematics*, Boca Raton, Fl.: Chapman & Hall/CRC, 2000.

[Jam] James, I.M., ed., *History of Topology*, Amsterdam: Elsevier Science B.V., 1999.

[Jon] Jones, V.F.R., "A polynomial invariant for knots via von Neumann algebras" *Bulletin of the American Mathematical Society* **12** (1985), 103–112.

[Kak] Kakutani, S., "A Generalization of Brouwer's Fixed Point Theorem" *Duke Mathematical Journal* **8** (1941), 457–459.

[Kha1] Khalimsky, E., "Topological structures in computer science" *Journal of Applied Mathematics and Stochastic Analysis* **1** (1987), no. 1, 25–40.

[Kha2] Khalimsky, E., Kopperman, R., and Meyer, P.R., "Computer graphics and connected topologies on finite ordered sets" *Topology and its Applications* **36** (1990), 1–17.

[Kha3] Khalimsky, E., Kopperman, R., and Meyer, P.R., "Boundaries in Digital Planes" *Journal of Applied Mathematics and Stochastic Analysis* **3** (1990), no. 1, 27–55.

[Kis] Kiselman, C.O., "Digital Jordan Curve Theorems" In: Borgefors, G., Nyström, I., and Sanniti di Baja, G., eds., *Discrete Geometry for Computer Imagery*, DGCI 2000 Proceedings, Lecture Notes in Computer Science, Volume 1953, New York: Springer–Verlag, 2000, 46–56.

[Kon] Kong, T.Y., Kopperman, R., and Meyer, P.R., "A topological approach to digital topology" *American Mathematical Monthly* **98** (1991), no. 10, 901–917.

[Kov] Kovalevsky, V.A., "Finite topology as applied to image analysis" *Computer Vision, Graphics, and Image Processing* **46** (1989), 141–161.

[Kur] Kuratowski, K., "Sur le problème des courbes gauches en topologie" *Fundamenta Mathematicae* **15** (1930), 271–283.

[Lat] Latombe, J.-C., *Robot Motion Planning*, Boston: Kluwer Academic Press, 1991.

[Leh] Lehoucq, R., Lachièze-Rey, M., and Luminet, J.P., "Cosmic Crystallography" *Astronomy and Astrophysics* **313** (1996), 339–346.

[Lis] Listing, J.B. "Vorstudien zur Topologie" *Göttingen Studien* (1847), 811-875

[Liv] Livingston, C., *Knot Theory*, Carus Mathematical Monographs 24, Washington D.C.: Mathematical Association of America, 1993.

[Lor] Lorenz, E.N., "Deterministic nonperiodic flow" *Journal of Atmospheric Sciences* **20** (1963), 130–141.

[Mae] Maehara, R., "The Jordan curve theorem via the Brouwer fixed point theorem" *American Mathematical Monthly* **91** (1984), no. 10, 641–643.

[Mak] Makarychev, Y., "A Short Proof Of Kuratowski's Graph Planarity Criterion" *Journal of Graph Theory* **25** (1997), 129–131.

[Man] Manheim, J.H., *The Genesis of Point Set Topology*, New York: The Macmillan Company, 1964.

[Mar] Markov, A.A.,"Insolubility of the problem of homeomorphy" In: Todd, J.A., ed., *Proceedings of the International Congress of Mathematicians*, Cambridge: Cambridge University Press, 1958, 300–306.

[Mas] Massey, W.S., *A Basic Course in Algebraic Topology*, New York: Springer–Verlag, 1991.

[May] May, R.M., "Simple mathematical models with very complicated dynamics" *Nature* **261** (1976), 459–467.

[Moi] Moise, E.E., *Geometric Topology in Dimensions 2 and 3*, New York: Springer–Verlag, 1977.

[Mon] Monastyrsky, M., *Riemann, Topology, and Physics*, translated by Cook, R., King, J., and King, V., Boston: Birkhäuser, 1999.

[Mun] Munkres, J.R., *Topology*, Upper Saddle River, N.J.: Prentice Hall, Inc., 2000.

[NasC1] Nash, C., "Topology and physics—a historical essay" In: James, I.M., ed., *History of Topology*, Amsterdam: Elsevier Science B.V., 1999, 359–416.

[NasC2] Nash, C., and Sen, S., *Topology and Geometry for Physicists* New York: Academic Press, 1983.

[NasJ] Nash, J.F., "Equilibrium Points in n-Person Games" *Proceedings of the National Academy of Sciences* **36** (1950), 48–49.

[New] Newman, J.R., ed., "Leonhard Euler and the Koenigsberg Bridges" *Scientific American* **189** (1983), no. 1, 66–70. (This article primarily consists of a translation of Euler's original paper on the subject.)

[Now] Nowacki, W., "Die euklidischen, driedimensionalen, geschlossenen und offenen Raumformen" *Commentarii Mathematici Helvetici* **7** (1934), 81–93.

[Pet] Peterson, I., *Newton's Clock: Chaos in the Solar System*, New York: W.H. Freeman and Company, 1993.

[Ple] Pless, V., *Introduction to the Theory of Error-Correcting Codes*, New York: John Wiley & Sons, Inc., 1998.

[Rad] Radó, T., "Über den Begriff der Riemannsche Fläche" *Acta Mathematica Szeged* **2** (1925), 101–121.

[Rei1] Reidemeister, K., "Elementare Begründung der Knotentheorie" *Abhandlungen aus dem Mathematischen Seminar der Universität Hamburg* **5** (1927), 24–32.

[Rei2] Reidemeister, K., *Knotentheorie* Berlin: Springer, 1932; New York: Chelsea, 1948. (English translation: *Knot Theory* Moscow, Id.: BCS Associates, 1984.)

[Ren] Renz, J., *Qualitative Spatial Reasoning with Topological Information*, Lecture Notes in Computer Science, Volume 2293, New York: Springer–Verlag, 2002.

[Roh] Rohrer, R.A., *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems: Special Issue on Routing in Microelectronics* **CAD-2** (1983), no. 4.

[Ros] Rosenfeld, A., "Digital topology" *American Mathematical Monthly* **86** (1979), 621–630.

[Rou] Rouvray, D.H., "Predicting Chemistry from Topology" *Scientific American* **255** (1986), 40–47.

[Ryd] Ryden, K., ed., *Open GIS Implementation Specification for Geographic Information - Simple Feature Access - Part 1: Common Architecture*, Open Geospatial Consortium, Inc., 2005.

[San] Sankoff, D., and Kruskal, J., eds., *Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparison*, Stanford, Ca: CSLI Publications, 1999.

[Sar] Sarkaria, K.S., "The topological work of Henri Poincaré" In: James, I.M., ed., *History of Topology*, Amsterdam: Elsevier Science B.V., 1999, 123–168.

[Sch] Schneider, M., and Behr, T., "Topological Relationships Between Complex Spatial Objects" *ACM Transactions on Database Systems* **31** (2006), no. 1, 39–81.

[Sha] Shariff, A.R., Egenhofer, M., and Mark, D., "Natural-Language Spatial Relations Between Linear and Areal Objects: The Topology and Metric of English-Language Terms" *International Journal of Geographical Information Systems* **12** (1998), no. 3, 215–246.

[Shv] Shvarts, A.S., *Topology for Physicists*, translated by Levy, S., New York: Springer–Verlag, 1994.

[Sta1] Stadler, B.M.R., and Stadler, P.F., "The Topology of Evolutionary Biology" in Ciobanu, G., and Rozenberg, G., eds., *Modeling in Molecular Biology*, Natural Computing Series, New York: Springer–Verlag, 2004, 267–286.

[Sta2] Stadler, B.M.R., Stadler, P.F., Wagner, G., and Fontana, W., "The topology of the possible: Formal spaces underlying patterns of evolutionary change" *Journal of Theoretical Biology* **213** (2001), 241–274.

[Str] Strogatz, S.H., *Nonlinear Dynamics and Chaos: With Applications to Physics, Biology, Chemistry and Engineering*, Boulder, Co.: Westview Press, 2001.

[Sum] Sumners, D.L., ed., *New Scientific Applications of Geometry and Topology*, Providence, R.I.: American Mathematical Society, 1992.

[The] Theobald, D.M., "Topology revisited: representing spatial relations" *International Journal of Geographical Information Systems* **15** (2001), no. 8, 689–705.

[Tho] Thomassen, C., "Kuratowski's Theorem" *Journal of Graph Theory* **5** (1981), 225–241.

[Thu1] Thurston, W.P., and Levy, S., ed., *Three-Dimensional Geometry and Topology*, Princeton, N.J.: Princeton University Press, 1997.

[Thu2] Thurston, W.P., and Weeks, J.R., "The mathematics of three-dimensional manifolds" *Scientific American* **251** (1984), no. 1, 108–120.

[Tou] Touhey, P., "Yet another definition of chaos" *American Mathematical Monthly* **104** (1997), 411–413.

[Tri] Trinajstić, N., *Chemical Graph Theory*, Boca Raton, Fl.: CRC Press, 1992.

[Tur] Turan, P., "A Note of Welcome" *Journal of Graph Theory* **1** (1977), 7–9.

[Veb] Veblen, O., "Theory on plane curves in non-metrical analysis situs" *Transactions of the American Mathematical Society* **6** (1905), 83–98.

[Vel] Vellekoop, M., and Berglund, R., "On Intervals, Transitivity = Chaos" *American Mathematical Monthly* **101** (1994), 353–355.

[Vic] Vick, J. W., *Homology Theory: An Introduction to Algebraic Topology*, New York: Springer–Verlag, 1994.

[Von] Von Neumann, J., and Morgenstern, O., *Theory of Games and Economic Behavior*, Princeton, N.J.: Princeton University Press, 1944.

[Wee1] Weeks, J.R., "Reconstructing the global topology of the universe from the cosmic microwave background" *Classical and Quantum Gravity* **15** (1998), 2599–2604.

[Wee2] Weeks, J.R., *The Shape of Space*, New York: Marcel Dekker, Inc., 2002.

[Wie] Wiener, H., "Structural determination of paraffin boiling points" *Journal of the American Chemical Society* **69** (1947), 17–20.

[Win] Winfree, A.T., "Sudden Cardiac Death: A Problem in Topology" *Scientific American* **248** (1983), no. 5, 144–161.

Index